2021 年度学士論文

教師なし学習による音楽性を考慮した

自動ドラム採譜モデル

Unsupervised Automatic Drum Transcription

With Consideration Of Musicality

担当教授　黄 潤和

法政大学

情報科学部ディジタルメディア学科

黄研究室

18K1005　大石（おおいし）　芳弘（よしひろ）

目次

1. はじめに ..... - 1 -

2. 関連研究 ..... - 2 -

2.1. 教師なし学習を用いた研究 ..... - 2 -

2.2. 音楽性を考慮した研究 ..... - 2 -

3. 提案手法 ..... - 3 -

3.1. 問題設定 ..... - 4 -

3.1. エンコーダモデル ..... - 4 -

3.2.1. U-Net ..... - 5 -

3.2.2. Transformer ..... - 5 -

3.2.3. Sparsemax と Upsampling ..... - 7 -

3.3. デコーダモデル ..... - 8 -

3.4. 学習方法 ..... - 8 -

4. 実験 ..... - 9 -

4.1. 使用データセット ..... - 9 -

4.1.1. 学習データセット ..... - 9 -

4.1.2. 評価データセット ..... - 10 -

4.1.3. 推定音声のためのサンプル音源 ..... - 10 -

4.2. 実験詳細 ..... - 11 -

5. 実験結果と考察 ..... - 11 -

5.1. サンプリングレート・活性化関数の実験 ..... - 11 -

5.2. 先行教師なしモデルとの比較 ..... - 12 -

5.3. 本研究モデルの詳細分析 ..... - 14 -

5.3.1. 学習過程の分析 ..... - 14 -

5.3.2. 推定譜の分析 ..... - 15 -

6. おわりに ..... - 16 -

謝 辞..... - 17 -
文 献..... - 17 -

教師なし学習による音楽性を考慮した自動ドラム採譜モデル

Unsupervised Automatic Drum Transcription Model With Consideration of Musicality

大石　芳弘
Yoshihiro Oishi

法政大学情報科学部ディジタルメディア学科
E-mail: yoshihiro.oishi.9d@stu.hosei.ac.jp

Abstract

This paper describes an automatic drum transcription (ADT) system using unsupervised learning. Automatic drum transcription is the subtask of an automatic music transcription. This system converts from an audio source to drum instrument note onsets. There are several methods to estimate the music score, and I use machine learning as the learning method. In general, a supervised learning is the mainstream for this task, but few large and annotated datasets for drum transcription exist. Therefore, this system is built by unsupervised learning and has encoder-decoder model. It learns the model by self-learning. The model obtains the estimated score by putting the output of the encoder into score format. Recently, ADT systems considering musicality such as grammar, theory, and more are gathering attention. This is because simple estimation has limited accuracy. In this study, model has Transformer structure. Transformer can capture the musical characteristics, especially repeated structure, measure, phrase, and ABA structure etc. Through experimentation, this system can achieve the same level of accuracy as previous models both supervised and unsupervised models. And this study has succeeded in significantly speeding up the process compared to path unsupervised model.

1. はじめに

自動音楽採譜とは過去に録音された音響信号からその楽譜, つまり楽器のオンセット位置を推定することを目的としたタスクである. このタスクによって音楽制作・音楽教育・音楽学などの様々な分野での応用が可能となるため, 音楽タスク全般において非常に重要な研究テーマである. また, 音楽は様々な楽器によって構成されているが, その中でも特にドラムはリズムを形成・強調する重要な要素である. そのため, 自動音楽採譜のタスクの一つである自動ドラム採譜によって, リズムに関連した様々な音楽処理が可能となる. 本稿ではドラム音のみで構成されている音声信号から, キックドラムやスネアドラムなどの各ドラム構成要素のオンセット位置の推定を行う自動ドラム採譜モデルを作成する.

これまでの自動ドラム採譜の研究では, 非負値行列因子分解や隠れマルコフモデルが採用されており, 近年では機械学習や深層学習が主流となっている. 機械学習には大きく分けて教師あり学習と教師なし学習があり, 精度の面などの理由によって教師あり学習が採用されている. 自動ドラム採譜の教師あり学習には入力となるドラム音声信号と, 正解ラベルとなるすべての構成要素のオンセット位置の二つがデータセットとして必要となる. しかし, 現時点においてそのような自動ドラム採譜に用いることのできる大規模なアノテーション付きデータセットはあまり存在しない. そのため, 教師あり学習を採用する際には自らデータセットを作成する手間が必要となる. さらに, データセットの質はモデルの精度に直結するため, 作成したアノテーションの正確性を保証する必要がある. そこで, 本研究はドラムの音声信号のみで学習をすることが可能な教師なし学習による自動ドラム採譜システムを構築する. モデルにはエンコーダ・デコーダモデルを採用し, モデルの入出力を同じにするように自己学習を行うため, アノテーションとしてオンセット位置を必要とせずドラム音声信号のみで学習することが出来る. モデルのエンコーダでは畳み込みネットワークやリカレントネットワークによって, 入力であるドラム音声信号から各構成要素のオンセット位置を求める. モデルのデコーダは, エンコーダで求めたドラム推定楽譜のオンセット位置にドラム構成要素の単音であるサンプル音源を畳み込むことにより推定音声信号を求める.

さらに, 近年の自動ドラム採譜では音楽的側面を考慮して楽譜の推定を行う研究も注目されている. 従来の多くの研究では, 入力音声信号の特徴のみを学習することによってドラムのオンセット位置を推定している. しかし, ドラム音声の特徴のみでは推定したオンセット位置が音楽理論の上で妥当であるかを考慮していないため, 精度が頭打ちになってしまうことが多々ある. そこで, 本研究では繰り返し構造などの音楽的特徴を再現することができる Transformer をモデルに組み込んで学習を行う. また, Transformer は一般的に時系列を処理する RNN などのモデルより広範な構造を学習可能であり, 並列計算を高速化することができるため, これらの理由からも Transformer を採用する.

2. 関連研究

本章では，本研究の参考にした自動ドラム採譜の既存研究について述べる．これまでの自動ドラム採譜研究において用いられてきた代表的な手法として，非負値行列因子分解や隠れマルコフモデルがある．加えて，近年の自動ドラム採譜を含む音楽研究分野で頻繁に用いられている手法として機械学習や深層学習が挙げられる．畳み込みネットワークの CNN やリカレントネットワークの RNN などは，それぞれ時間 – 周波数領域・時系列を考慮することができるため高い精度でドラム譜を推定することができる[1].

2.1. 教師なし学習を用いた研究

Keunwoo ら[2]は教師なし学習によって自動ドラム採譜を実現した．教師なし学習を採用することによって，教師あり学習に比べてスケーラブルで一般化が可能となる．モデルにはエンコーダ・デコーダ構造を採用しており，オートエンコーダ型のネットワークを使用することによって教師なし学習を実現している．エンコーダではドラム音のみの入力音声を推定譜に変換しデコーダで推定譜から入力音声を復元した推定音声に変換している．モデルの学習には入力音声と推定音声の平均絶対誤差を使用した自己学習を行い，学習したモデルのエンコーダのみを使用することによってドラム譜を得る．また，ドラムのオンセットには要素・時間方向にスパース性が存在する．このスパース性を再現するために，エンコーダの活性化関数に Sparsemax を使用している．さらに，ドラム音の持続部分はドラムキットによって大きく異なり，入力音声と推定音声は同じドラムキットを使用しないため誤差が大きくなってしまう．そこで，入力音声と推定音声に対してメディアンフィルタによるオンセット強調を行うことによって，持続部分を削除し打点部分のみを残すことができる．これによって誤差を小さくすることと同時に，打点位置のみが残ることによって疑似的な譜面の誤差として扱うことができる．最終的な結果として，SMT 評価データセットにおいて 86.9%と高い精度を出しており，既存の様々な教師あり・なしモデルの精度を上回っている．また，実行時間や一般化可能性の面でも既存教師ありモデルより優れている．この研究により，自動ドラム採譜における教師なし学習の有用性やスパース性の考慮重要性を示した．

2.2. 音楽性を考慮した研究

近年では推定した楽譜の音楽性を考慮した研究も行われている．既存研究では入力音声の特徴から単純にドラムのオンセット位置を推定しているため，音楽文法やフレーズのパターン等の音楽的側面を考慮することが出来ずに精度が頭打ちになってしまう場合が多々ある．そこで，吉井ら[3]は言語モデルを用いることによって音楽性を確保した．言語モデルは推定したオンセット位置を確率的(数値的)に評価することができるため，推定譜が音楽的に妥当であるかを判定することができる．また，ほとんどの自動ドラム採譜研究では楽譜を

フレーム単位(秒)で推定しているが，実際の楽譜はテイタム単位(音符)で表現している．そこで，推定譜をテイタム単位で出力することによって音楽性と実用性を確保している．

また，近年自然言語処理に頻繁に用いられている Transformer は，音楽タスクにおいても非常によい結果を残している．自動作曲タスクでは，Huang ら[4]は Transformer によって音楽の繰り返し構造を再現している．音楽には ABA 形式や小節などメロディーやリズムを反復する構造があり，そのような繰り返し構造は音楽の重要な要素である．Transformer に組み込まれている自己注意機構によって小節やフレーズを単位とするパターンを認識し，先に述べたような繰り返し構造を再現することができる．さらに，吉井らの研究においても時系列特徴の抽出に Transformer を用いてモデルを構築しており，Transformer が自動ドラム採譜を含む音楽タスクにマッチしていることが分かる．

以上より，音楽性を考慮することによって推定譜の妥当性を確保し，モデルの精度を向上させることができる．

3. 提案手法

本研究では，教師なし学習による音楽性を考慮した自動ドラム採譜モデルを提案する．自動ドラム採譜において，教師あり学習に使用することのできる大規模なアノテーション付きデータセットはあまり存在しない．よって，モデルの学習方法には教師なし学習を採用する．教師なし学習によってモデルの学習を行うため，モデルにはエンコーダ・デコーダ構造を使用する．エンコーダによって入力音声を推定譜へ変換し，デコーダで推定譜から入力音声を模した推定音声を生成する．また，音楽的側面を考慮するために繰り返し構造を再現することができる Transformer をモデルに組み込む．本章では，初めにモデルの入出力などの本研究の問題設定(2.1 節)，続いて教師なしモデルのエンコーダ部分の詳細(2.2 節)・デコーダ部分の詳細(2.3 節)について述べる．また，教師なしモデルの学習方法(2.4 節)についても述べる．提案モデルの全体図は図 1 のようになる．

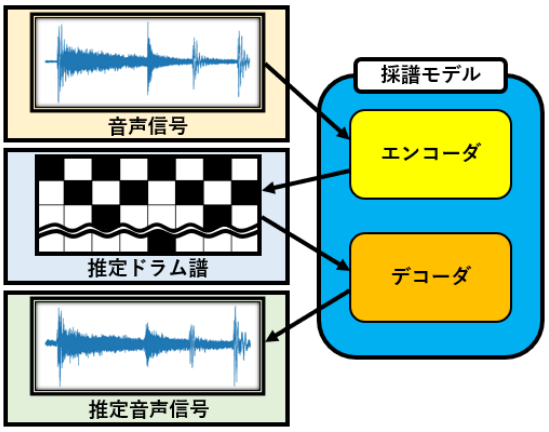

図 1．提案モデルの全体図

3.1. 問題設定

本研究では，ドラム音のみで構成された音声信号$X \in \mathbb{R}^{1\times T}$からドラム譜$Y \in \mathbb{R}^{K\times T}$を推定する自動ドラム採譜モデルを作成する．ここで，K は本研究で設定しているドラムの構成要素数(ハイハットやスネアドラム等)，T は時間フレーム数を表す．今回，ドラムの構成要素として考慮するものを表 1 に示す．表 1 より，本研究の構成要素数は$K = 16$となる．なお，ドラム構成要素のクラスやサブクラスなどの分類方法は[5]を参考にする．

表 1．使用したドラム構成要素一覧

Class	Subclass	Description
KD	KD	kick drum
SD	SD	snare drum
HH	CHH	closed hi-hat
	PHH	pedal hi-hat
	OHH	open hi-hat
TT	HIT	high tom
	MHT	high-mid tom
	HFT	high floor tom
CY	RDC	ride cymbal
	RDB	ride cymbal bell
	CRC	crash cymbal
	CHC	china cymbal
	SPC	splash cymbal
OT	TMB	tambourine
	CVS	claves
	CB	cowbell

3.1. エンコーダモデル

エンコーダでは，ドラム入力音声Xから推定譜Yを出力する．この時，入力信号Xは正規化した音声信号を使用し，推定譜Yは[0, 1]の値をとる配列であり数値の位置がドラムオンセットの位置，数値の大きさがベロシティ(音量)を表している．初めに畳み込みネットワークによって入力音声の特徴を抽出するために U-Net を使用する．その後，時系列を考慮するリカレント層として Transformer を使用する．推定譜Yを得るための活性化関数には，Keunwoo らを参考にしてドラム音声のスパース性を考慮するために Sparsemax を使用する．エンコーダの全体構造は図 2 のようになる．

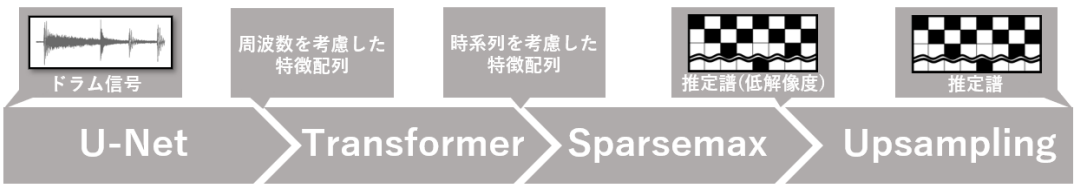

図 2．エンコーダの流れ

3.2.1. U-Net

エンコーダでは初めに，入力音声Xの時間・周波数的特徴を畳み込みネットワークによって抽出する．畳み込みを用いたモデルとして，画像セグメンテーションによく利用されている U-Net を使用する．U-Net の構造は図 3 のようになる．

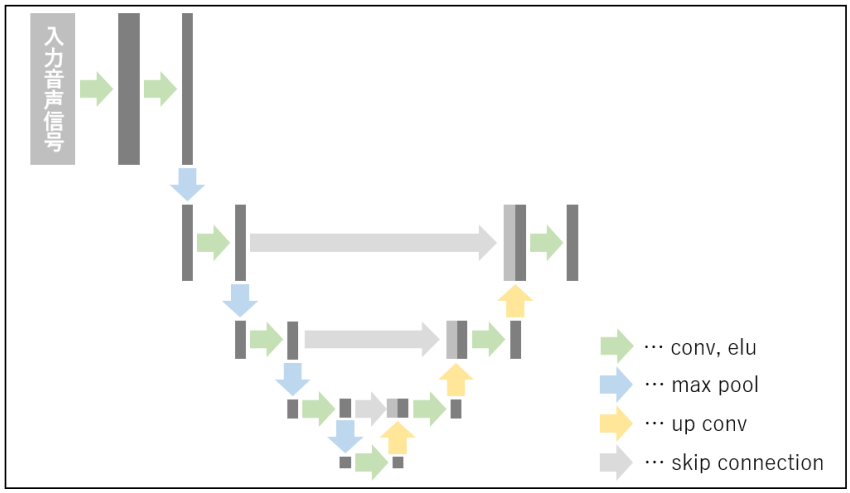

図 3．U-Net の構造

U-Net の左側では畳み込みを行っており，1 次元の畳み込みと elu 関数，そしてマックスプーリングによって構成されている．右側では先の畳み込みの処理とは反対の処理を行う逆畳み込みと，畳み込みの際に保持していた情報を同じ深さの特徴に加える skip connection によって構成されている．逆畳み込みは線形補間によって実現することができる．この時，畳み込み層と逆畳み込み層の入出力の変形スケールは同じであり，それぞれの深さの非対称性によって U-Net の出力は$r_1 \in \mathbb{R}^{D_F \times T/C}$と入力音声$X$よりも短くなる．ここで$D_F$は潜在ベクトルの次元数であり，畳み込み層のスケール変化をs，層の深さの差をdとしたとき$C = s^d$となる．よって，推定譜の解像度(サンプリングレート)は入力音声よりも低くなってしまい，その分細かいフレーム区分の推定を行うことが出来なくなる．

3.2.2. Transformer

U-Net では入力音声から時間・周波数の特徴を取り出した．しかし，入力音声は時系列データであるため，時系列の特徴も取り出した方がよい．音声処理や自然言語処理の際に，時系列の特徴を取り出すことに使用される代表的なモデルとして RNN や LSTM が挙げられる．また近年の研究において，自然言語処理などで頻繁に用いられる Transformer は音楽タスクにおいて非常に良好な結果を残すことが分かっている．特に自動作曲タスクにおいて，2.2 節で述べたように Transformer は音楽の繰り返し構造を再現でき，音楽の特徴を捉える

ことができるモデルであることが分かる．また，Transformer は RNN 等に比べてより広範な構造を学習することができ，並列計算も高速に行うことができるため推論時間を減少させることができる．よって，以上の理由より本研究においてもモデルに Transformer を組み込む．Transformer はエンコーダ・デコーダ構造であるが，ここでは時系列を考慮した特徴の抽出のみを目的としているため Transformer のエンコーダ部分のみを使用する．Transformer エンコーダの構造は図 4 のようになる．

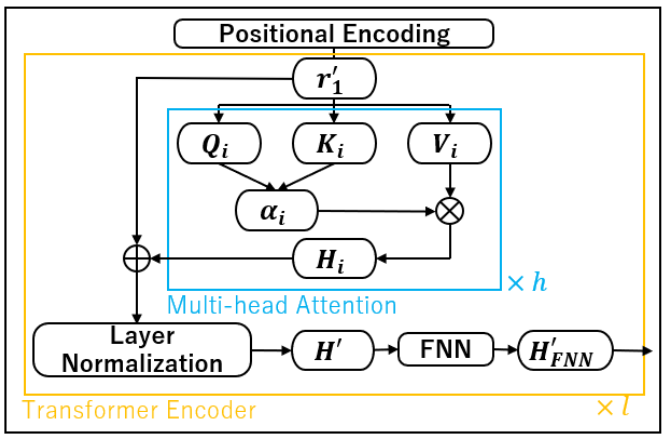

図 4．Transformer エンコーダの構造

Transformer では U-Net で抽出した特徴r_1に Positional Encoding を行い，位置情報を追加した新たな特徴$r'_1 \in \mathbb{R}^{K\times T/C}$を得る．その後，自己注意機構と Layer Normalization などによって時系列と音楽性を考慮した特徴$r_2 \in \mathbb{R}^{K\times T/C}$を出力とする．

Transformer エンコーダでは，自己注意機構によってそれぞれの入力位置が別の入力位置に注目したときの関連度を表すスコアを得ることができる．自己注意機構内のクエリベクトル$Q_i \in \mathbb{R}^{D_K\times T/C}$，キーベクトル$K_i \in \mathbb{R}^{D_K\times T/C}$，バリューベクトル$V_i \in \mathbb{R}^{D_K\times T/C}$はそれぞれ以下のように$r'_1$の全結合処理によって得られる．

$$Q_i = W_i^Q r'_1 + b_i^Q \tag{1}$$
$$K_i = W_i^K r'_1 + b_i^K \tag{2}$$
$$V_i = W_i^V r'_1 + b_i^V \tag{3}$$

このときの全結合処理の重みとバイアスは$W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{D_K\times D_F}, b_i^Q, b_i^K, b_i^V \in \mathbb{R}^{D_K\times T/C}$であり，Multi-head Attention の head 数を$h \in \mathbb{N}$としたとき$D_k = \frac{D_F}{h}$である．注意重み$\alpha_i \in \mathbb{R}^{T/C\times T/C}$は，入力である特徴$r'_1$自身の関連度を表す正規化された行列であり，以下の計算によって得ることができる．

$$e_{i,n,n'} = \frac{q_{i,n}^T k_{i,n'}}{\sqrt{D_k}} \tag{4}$$

$$\alpha_{i,n,n'} = softmax\left(e_{i,n,n'}\right) \tag{5}$$

このとき，$q_{i,n}$はQ_iの n 番目の列であり，nとn'はそれぞれQ_iとK_iの時間方向のインデックスである．その後，バリューベクトルV_iと注意重みα_iの内積を潜在表現ベクトル方向に結合することによって，潜在ベクトル表現$H \in \mathbb{R}^{D_F \times T/C}$を得ることができる．$H$に入力$r_1'$を加え，Layer Normalization の処理を行うことで$H' \in \mathbb{R}^{D_F \times T/C}$を得る．その後，以下の計算によって 2 層の全結合層と 1 つの Relu 活性化関数によって 1 つのレイヤーの Transformer エンコーダ層の出力$H'_{FFN} \in \mathbb{R}^{D_F \times T/C}$を得る．

$$H'_{FFN} = W_2^{H'} \max\left(0, W_1^{H'} H' + b_1^{H'}\right) + b_2^{H'} \tag{6}$$

このとき，$W_1^{H'} \in \mathbb{R}^{D_{FFN} \times D_F}, W_2^{H'} \in \mathbb{R}^{D_F \times D_{FFN}}, b_1^{H'} \in \mathbb{R}^{D_{FFN} \times T/C}, b_1^{H'} \in \mathbb{R}^{D_F \times T/C}$である．その後，$H'_{FFN}$を次の Transformer エンコーダレイヤーへの入力とし，l回上記の処理を繰り返す．最後に全結合層によって潜在ベクトル表現の次元数D_Fを設定しているドラムの要素数Kに変形した時系列を考慮した特徴$r_2 \in \mathbb{R}^{K \times T/C}$を得る．

3.2.3. Sparsemax と Upsampling

従来の自動ドラム採譜では，推定譜をシグモイド関数やソフトマックス関数などの深層学習によく用いられる活性化関数によって求めている．しかし，想定するドラムのオンセットは時間フレームで連続することはなく，ドラム構成要素において一度に鳴らされる数は限りがある．つまり，ドラム音源には時間方向とドラム構成要素方向の両方にスパース性が存在することになる．そのためスパース性を再現するために，先の一般的な活性化関数ではなく Sparsemax を採用する．Sparsemax の特徴として，ソフトマックス関数と同じように出力の和が常に 1 となり，ソフトマックス関数とは違って確率として低いものは 0 になる特徴を持っている．Sparsemax の定義として，

$$\triangle^{K-1} \coloneqq \left\{p \in \mathbb{R}^K \middle| 1^T p = 1,\ p \geq 0\right\} \tag{7}$$

$$Sparsemax(z) \coloneqq argmin||p - z||^2_{p \in \triangle^{K-1}} \tag{8}$$

となる．このとき，n 次元ベクトル z の次元数を K とする．よって，時間方向とドラム要素方向のスパース性を確保するために，Transformer の出力r_2に対して時間方向と要素方向の別々に Sparsemax を適用させる．それぞれの方向に対しての出力を得た後，それらの配列を要素ごとに掛け合わせることによって推定譜を求める．また，ドラム信号は時間方向にスパース性が存在すると定義したが，ドラム構成要素の中でハイハットはその他の要素に比べ

てオンセット頻度が非常に多い．そこで，本研究ではハイハットにのみスパース性を考慮しないようにするために，ソフトマックスをハイハットにのみ適用し，その他の要素にはSparsemax を適用した場合の実験も行う．

また，U-Net の非対称性によって Sparsemax の出力長は入力音声よりも短くなっている．そこで，入力長と同じにするために 0 挿入のアップサンプリングによって推定譜Yを求める．

3.3. デコーダモデル

デコーダでは，推定譜Yから入力音声Xを復元した推定音声$\hat{X}$を出力する．推定音声$\hat{X}$の合成には，推定譜Yと各ドラム構成要素のサンプル音源を使用する．サンプル音源は 1 秒間の正規化された単体音源である．推定譜Yでオンセット判定されたすべての箇所に巡回畳み込みによって各サンプル音源を当てはめ，ドラム構成要素ごとの推定音声$\hat{X}_k$を求める．しかし，単純な巡回畳み込みの計算量は$O(N^2)$であるため，実行時間とメモリに負担がかかってしまう．そこで，フーリエ変換による巡回畳み込みの高速化によって計算量を$O(N \log N)$にすることができる．高速化した巡回畳み込みによって各ドラム構成要素の推定音声$\hat{X}_k$を求める式は以下のようになる．

$$Y_k * x_k = \mathcal{F}^{-1}\mathcal{F}(Y_k * x_k) = \mathcal{F}^{-1}\Big(\big(\mathcal{F}(Y_k)\big)\big(\mathcal{F}(x_k)\big)\Big) \tag{9}$$

この時，Y_kは一つのドラム構成要素の推定譜，x_kはそのドラム構成要素のサンプル音源の信号，$\mathcal{F}$はフーリエ変換を表しており最後のフーリエ変換後の乗算はアダマール積である．

それぞれのドラム構成要素の推定音声$\hat{X}_k$を計算後，すべての推定音声の総和によって最終出力である推定音声$\hat{X}$を求めることができる．なお，このデコーダ部分はモデルの学習時にのみ使用される．

3.4. 学習方法

本研究のモデルはエンコーダ・デコーダ構造であり，入力音声を模した推定信号を出力する．この時，入力音声 X と推定音声$\hat{X}$の平均絶対誤差を損失とし，この損失を減少させるようにモデルのパラメータを調整していく．誤差を計算する際には，それぞれの音声を周波数領域に拡張するためにフーリエ変換を行う．本研究では一般的なフーリエ変換は利用せず，定 Q 変換を採用する．定 Q 変換は一般的に音楽を対象に，音高・コード・メロディなどを分析する際に使用される．これは，定 Q 変換では中心周波数に応じて時間窓のサンプル数を変更し周波数分解能を変更することによって，周波数の低域と高域で分解能を一定にすることができるためである．離散時間領域信号を$x(n)$とし，定 Q 変換のスペクトル系列は

$$X^{CQ}(k,\ n) = \sum_{j=n-\left\lfloor \frac{N_k}{2} \right\rfloor}^{n+\left\lfloor \frac{N_k}{2} \right\rfloor} x(j) a_k^*\left(j - n + \frac{N_k}{2}\right) \tag{10}$$

となる．ここで，$k = 0,\ 1,\ \ldots,\ K-1$は定 Q 変換の周波数ビン，$n = 0,\ 1,\ \ldots,\ N-1$は時間フレームを表す．また，$a_k^*(n)$は$a_k(n)$の複素共役であり，

$$a_k(n) = \frac{1}{N_k} w\left(\frac{n}{N_k}\right) \exp\left[-\frac{i2\pi nQ}{N_k}\right] \tag{11}$$

である．これによって中心周波数が高いときは窓幅を小さくし，低いときは窓幅を大きくすることで分解能を同じにすることができる．今回の定 Q 変換の周波数帯として 32.07Hz(C1)から 8kHz(C8)までの帯域を 1 オクターブあたり 12 個のビン(C~B までそれぞれ 1 ビン)で解析する．以上の定 Q 変換の処理を入力音声 X と推定音声$\hat{X}$にそれぞれ行い，その平均絶対誤差を損失として利用して学習を行う．

4. 実験

本章では，実験を行った際の環境設定について述べる．4.1 節では学習と評価に使用したデータセットと推定音声の合成に利用したサンプル音源について述べる．また，モデルのパラメータや評価の際に適用したピークピッキング等の実験詳細を 4.2 節で述べる．

4.1. 使用データセット

4.1.1. 学習データセット

モデルの学習に使用するデータセットとして，Groove MIDI Dataset と ADT 用に作成された合成音声データセット[6]を使用する．

1 つ目の Groove MIDI Dataset は約 13.6 時間分の 10 人のプロドラマーによって演奏されたデータセットである．それぞれのドラマーは電子ドラムを用いて複数のスタイルで即興演奏を行い，ジャンル数や演奏の質を保証した高品質なデータセットとなっている．

2 つ目の合成音声データセットは，自動ドラム採譜用に作成された約 259 時間の西洋音楽の MIDI 音源である．データセットは既存の MIDI 音源を使用しており，様々なジャンルを網羅している．また，このデータセットを作成するうえで各ドラム構成要素のオンセット分布の調整を行っている．一般的なドラム音源ではキックドラムやスネアドラム等の主要な

要素のみで構成されており，タムタムやカウベル等のマイナー要素はあまり使用されない．そのため，出現頻度の少ない要素はオンセット予測の際に無視されてしまうことが多々ある．そのようなマイナー要素を正しく予測するために，マイナー要素のオンセット頻度を増やすなどの調整を行っている．また，そのほかにも極端に演奏時間が長い，もしくは短いデータを削除する等の学習に不適切なデータの処理も行っている．

データセット内の各音声信号から，ランダムに 2 秒間の音声を取り出しそれを学習に使用する．なお，学習が 10epoch 進むたびにこのランダムピックの処理を行いモデルの入力データセットを作り直す．また．過去の自動ドラム採譜研究では実際に演奏されたデータセットと合成音声データセットを組み合わせるとよい結果が得られることが分かっている．合成音声は実際には演奏されないパターンを多く含んでいるため，データとしての幅を広げることができるためである．Groove MIDI Dataset は実際のドラム音源ではないが演奏自体は行われているため，本研究ではこの 2 つのデータセットを学習に使用する．

4.1.2. 評価データセット

自動ドラム採譜は，いくつかのタスクに分類することができる．本研究はドラムのみで構成された音声信号からオンセット推定を行うため，それに準じた 2 つのタスクを評価する．

1 つ目はキックドラム・スネアドラム・ハイハットの 3 種類のドラムメジャー要素でのみ構成された音声信号のオンセットを推定する Drum Transcription of Drum-only Recordings(DTD)を評価する．このタスクのデータセットには，自動ドラム採譜によく用いられる IDMT-SMT-Drums(SMT)を使用する．これは約 130 分の上記 3 種類の要素のみで構成されたドラム音源であり，ドラムキットには実世界のアコースティックドラムセットやシンセサイザーを利用している．また，それぞれの要素に対して手動でアノテーションが付けられている．

もう 1 つのタスクは，すべてのドラム構成要素を含んだ音声信号のオンセット推定を行う Drum Transcription in the Presence of Additional Percussion(DTP)である．こちらには，Medley-DB Drums(MDB)を使用する．こちらは合計 20 分のデータセットであり，SMT と同様に DTP の評価に用いられている．

これら 2 つの評価データセットに対して前処理などの微調整は行わず，全データを評価にのみ使用する．

4.1.3. 推定音声のためのサンプル音源

推定音声合成のサンプル音源には，2.1 節の Keunwoo らが使用していた音源がインターネット上に公開されているためそちらを使用する．サンプル音源は構成要素が鳴り始めてから 1 秒分を使用する．また，本研究の推定要素としてチャイナシンバルやカウベルなどを新たに追加しているため，それらの音源は DTM ソフトに収録されている音源を使用する．ただし，

音源としての質を確保するために，本研究では GM 規格に従っている音源のみをサンプル音源として使用する．

4.2. 実験詳細

初めに，モデルのパラメータについて述べる．U-Net の畳み込みの深さは 10，逆畳み込みの深さは 6 で構成する．1 つ目の畳み込みは 1 次元の(channel，kernel，stride) = (128，3，1)のものを使用し，その他の畳み込みには(50，3，1)のものを使用する．Transformer のレイヤー数は 5，Multi-head Attention の head 数は 5 とする．また，Layer Normalization 後の全結合処理のD_{FFN}は 200 とする．すべての重みの初期値には He の初期値を使用し，最適化手法には Adam を採用している．学習率は 0.004 とし，100epoch の学習を行う．

次に，評価方法について述べる．評価には推定譜にピークピッキングを行ったうえで精度を求める．ピークピッキングには信号$x[n]$が以下の 3 つの条件を満たす場合，サンプル n をピークとして検出する．

$$x[n] = \max\left(x\left[n - \mathrm{m_{pre}} : n + m_{post}\right]\right) \tag{12}$$

$$x[n] \geq mean\left(x\left[n - a_{re} : n + a_{post}\right]\right) + delta \tag{13}$$

$$n - n_{pre} > wait \tag{14}$$

ここで，m_{pre}とm_{post}は最大値を計算する n の前後のサンプル数であり，a_{pre}とa_{post}は英金値を計算する n の前後のサンプル数である．また，$delta$は平均計算後の閾値であり，n_{pre}は 1 つ前にピークとして検出されたサンプル，$wait$はその直前のサンプルからの間隔である．本研究では$m_{pre} = m_{post} = SR/20$，$a_{pre} = a_{post} = SR/10$，$delta = 0.25$，$wait = SR/16$としてピークピッキングを行う．また，ピークピッキング後の推定譜と正解ラベルのオンセット位置の許容誤差は 0.05s とし，真陽性の組み合わせが最も多くなるように計算する．評価値には F 値を用いる．F 値は Precision と Recall の調和平均によって求めることができる．

5. 実験結果と考察

本章では，3 章で述べた時間方向の活性化関数を Sparsemax にしたモデルとハイハットにソフトマックスを適用したモデルの精度実験を行い，先行研究との比較及び考察を行う．また，SMT と MDB に対して採譜と評価を行った際の結果を示す．

5.1. サンプリングレート・活性化関数の実験

初めに，学習データセットを読み込む際のサンプリングレートを 16000Hz にしたモデルと 8000Hz にした時の Sparsemax のモデル，さらに学習データセットを 8000Hz で読み込み

ハイハットの時間方向にのみソフトマックス関数を適用させたモデルの比較実験を行う．図 5 にそれぞれの評価結果を載せる．

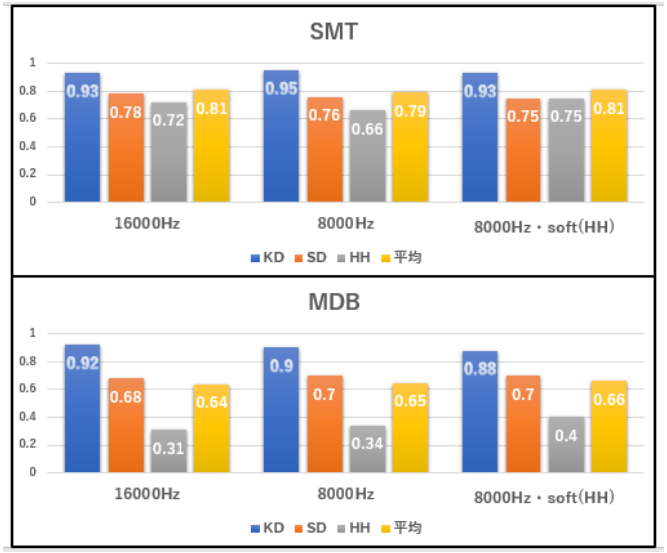

図 5．各モデルの F 値

16000Hz で読み込んだモデルと 8000Hz で読み込んだモデルの結果を比較すると，どちらの評価データに関してもあまり F 値に差がないことが分かる．つまり，自動ドラム採譜においてサンプリングレートは重要な要素ではないといえる．高いサンプリングレートほど精度が良くなるというわけではないため，特徴が失われない程度のサンプリングレートにまで下げることによってモデルへの入力長が短くなり推論時間を短縮することができる．しかし，U-Net のマックスプーリングによるダウンサンプリングや定 Q 変換のナイキスト周波数に影響が出てしまうため，適切なサンプリング周波数にする必要がある．

また，先ほどの 8000Hz のモデルと時間方向のハイハットにのみソフトマックス関数を適用させたモデルの比較を行う．結果を比べてみると，SMT と MDB のどちらにおいてもハイハットに Sparsemax を適用したものより，ソフトマックスを適用させた方がハイハットの F 値が向上していることが分かる．Keunwoo らの実験によって，活性化関数を全てソフトマックスにしたモデルはよい精度にならないことが分かっている．つまり，ハイハットにはその他の要素に比べてスパース性を考慮する必要がないことが分かる．これは，実際のオンセット分布においてもハイハットは最もオンセット回数が多いため，理にかなっている手法である．

5.2. 先行教師なしモデルとの比較

本節では，本研究のモデルと参考にした Keunwoo らの教師なしモデルの精度比較を行う．本研究のモデルには，5.1 節で紹介したソフトマックスを適用したモデルを比較に使用す

る．図 6 に先行研究と本研究の結果を示す．

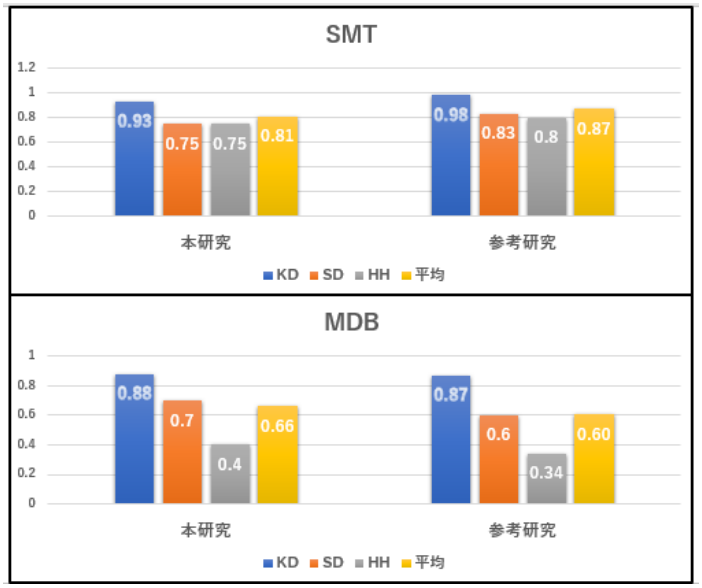

図 6．本研究と先行研究の精度比較

結果より SMT の精度，つまり DTD タスクにおいては先行研究のモデルが優れていた．しかし，MDB の精度はスネアドラムとハイハットの値が本研究の方が勝っており，平均 F 値も本研究が上回っている．よって，本研究は先行研究と比較して DTP タスクにおいて優れているといえる．SMT の精度が優れなかった原因として，モデルの構造が単純であるため要素間の特徴の違いを捉えきれていないことが挙げられる．また先行研究では GRU を要素方向に適用することによって要素方向の特徴をより抽出している．そのため，要素方向に特徴を抽出する新たなモデルを組み込むことで精度向上を行うことができる．一方で，MDB の結果が上回った要因として，Transformer と使用データセットが考えられる．先も述べたように，先行研究ではドラム構成要素方向を考慮するために GRU を適用していた．しかし，本研究は要素方向に適用させたネットワークは無いため，Transformer の自己注意機構が要素方向の特徴も捉えていることが分かる．これによって，Transformer は自動ドラム採譜のモデルに適しており，自己注意機構によって高い精度を得ることができると分かる．また，使用した合成データセットのオンセット分布調整によって，ハイハットやスネアドラムをシンバルやタムタムに誤判定してしまうケース，またはその逆のケースを防いでいると考えられる．

本研究は，推論時間の面においても先行研究を大幅に改善した．先行研究では学習に約 5 日かかっており，本研究ではわずか半日程度であった．先行研究の SMT の精度も本研究と同程度に至るまで約 1 日かかっているため，SMT と MDB のどちらにおいても推論時間

を短縮することが出来ている．これは，Transformer の並列計算の高速化によって実現しており，Transformer を用いることによって学習速度を格段に上昇させることができる．

よって，本研究は精度の面において，特に DTP タスクの精度は既存教師なしモデルを上回る結果を得た．また，学習時間の面でも大幅な時間短縮を行うことが出来た．以上により，Transformer が自動ドラム採譜に適応可能であることが分かる．

5.3. 本研究モデルの詳細分析

本節では，本研究の提案モデルの F 値のみだけでなく，それ以外の結果についても詳細分析を行う．

5.3.1. 学習過程の分析

5.1 節と 5.2 節では最も F 値が良かった epoch の結果を取り出している．ここでは本研究の 2 つのモデルの学習過程を分析する．図 7 にそれぞれのモデルの F 値推移を示す(左図：Sparsemax モデル，右図：ソフトマックスモデル，上図：SMT，下図：MDB).

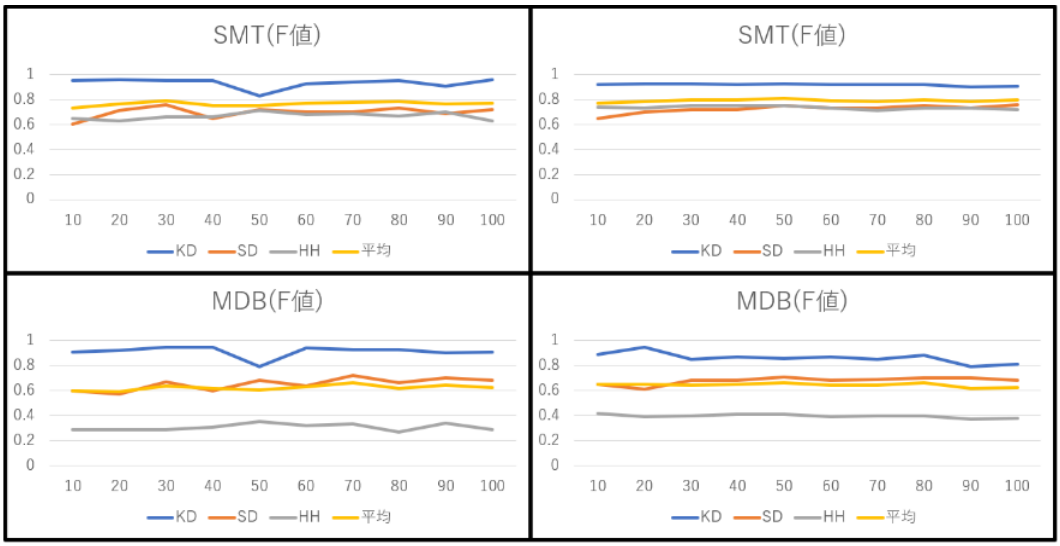

図 7. 学習推移図

上図より，どのモデルの要素に関しても 10epoch 目には最良に近い精度を出していた．つまり，本研究のモデルは特徴抽出を高速に行うことができることが分かる．加えて，Transformer が短時間で特徴を捉えることができるということが分かる．しかし，ほとんどの要素において 30epoch 付近から精度はほとんど変化しておらず，学習が収束しているためこれ以上の学習による精度向上は見込めない．これは，モデルの構造が単純であるため抽出できる特徴も

少ないからである．U-Net の畳み込みや Transformer のレイヤー数などを増やすことによって，より複雑な特徴を抽出することができ，精度向上につながる．

5.3.2. 推定譜の分析

ここでは，推定したオンセット位置とピークピッキングした推定譜，正解オンセット位置を比較，分析を行う．ここでは，先行研究よりもよい結果を得たハイハットにソフトマックスを適用したモデルの推定譜について分析を行う．

初めに，SMT の推定したオンセット位置について分析する．図 8 にそれぞれのオンセット図を示す(左列：キックドラム，中列：スネアドラム，右列：ハイハット，上段：推定譜，中段：ピークピッキング後の推定譜，下段：正解推定譜)．

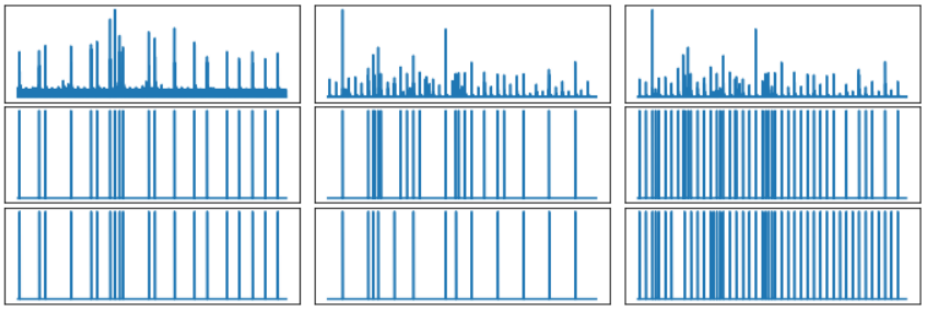

図 8 ． SMT の推定譜と正解譜

上図より，キックドラムはピークピッキング前後に関わらず高精度に推定することが出来ている．また，スネアドラムは正解オンセット位置にはほとんど推定することが出来ているが，余計なオンセットが少し見受けられる．ハイハットに関しては，後半に推定できていない箇所があることが分かる．しかし，ピークピッキング前の推定譜では小さいピークとして表れていることが分かる． これは，活性化関数をソフトマックスにしたことによって，Sparsemax に比べて一つ一つのピークの値が小さくなってしまうことが原因である．この解決策として，ピークピッキングの処理を要素ごとに変更するとよい．スネアドラムではよりスパース性を表現するためにピークの間隔や閾値を上げ，ハイハットには逆にピークの間隔や閾値を下げる．これによって，より要素間のスパース性の違いを表現することができる．

次に，MDB の推定オンセットについて分析を行う．図 9 に図 8 と同様の形式の図を示す．図より，キックドラムの精度は SMT より少し劣るもののよい結果であるが，スネアドラムとハイハットはあまり芳しい結果にはならなかった．スネアドラムとハイハットに似ているタムタムやシンバルを含んでいるが，それらの誤判定ではない誤りが非常に目立った．これは，学習データが 2 秒間の音声信号のみで学習しているためと考えられる．SMT は 20 秒

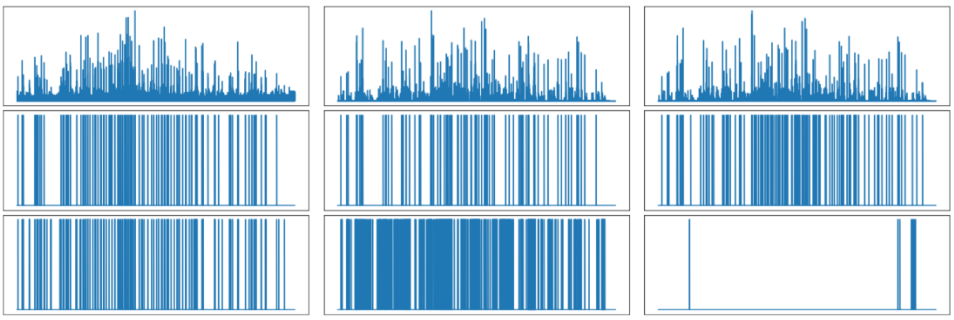

図 9. MDB の推定譜と正解譜

前後の音声信号であるが，MDB は約 50 秒前後となっているため長期的な音声信号に対応できていない．よって，より長い入力長にすることによって MDB のような長い音声に対応することができる．また，上図のハイハットのようにオンセットが無いところをオンセットとして判定している箇所がその他の楽曲に対しても非常に多かった．本研究の学習データセットを作成する際に無音が多いデータは削除していたため，無音状態を学習することができていないことも原因である．

MDB ではあまり確認することが出来なかったが，SMT では Transformer の繰り返し構造再現性の恩恵を受けることが出来ている．どの要素に関しても一定間隔でピークが出現しており，楽曲のリズムを捉えることが出来ている．しかし，リズムは最小の繰り返し単位であり，フレーズ等の繰り返し構造はあまり表現出来ていない．この問題に対しても先に述べたように，学習データセットの長さをより長くすることによってリズムだけでなくフレーズ等の繰り返し構造を入力することができるため，それらを学習することが出来る．

6. おわりに

本研究によって，Transformer は特徴抽出や推論時間の面から自動ドラム採譜に応用することができると判明した．また，ハイハットに対するスパース性の不必要性を明らかにすることができた．モデルの結果として，学習時間を既存の教師なしモデルに対して大幅に短縮することができ，DTP のタスクではその教師なしモデルの精度を上回る結果を残した．これにより，教師なし学習のボトルネックとなる推論時間を解決したので，必要なデータセットの量やモデルの複雑性等の問題を解決することによって，より教師なし学習の一般化可能性の恩恵を受けることができる．

今回は，推論時間の短縮や GPU のメモリ不足問題のためにモデル構造を簡潔にする必要があった．しかし，モデルを複雑にすることによってより特徴を捉えることができ，より良い精度を得ることができる．また，メモリの問題よりモデルへの音声の入力長を 2 秒にする必要があった．BPM が 120 の音楽でさえ 2 秒では 1 小節分の音声となってしまうため，Transformer の繰り返し構造再現性を十分に活用できていない．4 秒やそれ以上にすること

によってより Transformer の繰り返し構造再現性を生かすことができ，長期的な音声信号にも対応することができる．本研究ではメモリ不足のために断念した実験が多数あるため，より性能のよい GPU 環境で実験を行う，もしくはメモリ使用量を抑えるアルゴリズムによって更なる精度向上を望むことができる．

謝 辞

本研究を進めるに当たり，指導教官の黄教授及び TA の斎藤さんからは非常に沢山のご指導をして頂きました．深く感謝申し上げます．

文 献

[1] Wu, C.-W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Miiller, M. and Lerch, A. “A Review of Automatic Drum Transcription” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume 26, Issue 9, pp.1457-1483, September 2018.

[2] Keunwoo Choi, Kyunghyun Cho. “Deep Unsupervised Drum Transcription” Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR, Delft, TheNetherlands, pp.183-191, Jun 2019.

[3] 石塚崚斗， 錦見亮， 中村栄太， 吉井和佳， “大局的構造に基づく正則化を用いた自己注意機構付き深層ドラム採譜” 研究報告音楽情報科学(MUS), 2020-MUS-129, 3, pp.1-8, October 2020.

[4] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, lan Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, Douglas Eck. ”Music Transformer: Generating Music With Long-Term Structure” ICLR 2019, May 2019.

[5] C. Southall, C. Wu, A. Lerch, J. Hockman, “MDB Drums: An Annotated Subset of MedleyDB for Automatic Drum Transcription” Proceeding of the 18th International Society for Music Information Retrieval Conference, ISMIR, 2017.

[6] Richard Vogl, Gerhard Widmer, Peter Knees. ”Toward Multi-Instrument Drum Transcription” Proceedings of the 21th International Conference on Digital Audio Effects (DAFx18), September 2018.