修士論文

LSTMによる自動作曲システムの構築

2016年2月提出

指導教官　伊庭斉志　教授

情報理工学系研究科　電子情報学専攻

48-146440　姫野雅大

要旨

本研究では，自動作曲システムの構築及びその評価法の検討を行う．自動作曲システムはリカレントニューラルネットワークの一種である LSTM を用いて行う．また，LSTM を用いる前に，教師データに前処理を行うことによって次元の削減を行った．このシステムによって生成された音楽が創作物として適切であるかをを評価するため，類似度という指標を導入し，その妥当性を人間の感覚と比較することによって検討した．その結果，本研究によって使用した類似度は人間の感覚と強い相関があり，この類似度を用いることによって自動作曲システムによって生成された音楽は創作物として適切であると評価できるという結論に至った．

目 次

図 目 次

第 1章 序論

1.1 はじめに

コンピュータを用いた自動作曲は長年研究者を魅了してきた．[6] 人間の感性をコンピュータシステムに取り込むことは，コンピュータシステムの発展にとって必要不可欠という認識がある．[11] これまでに研究された自動作曲システムの一例として，対話型進化計算を用いた作曲システム [11]，歌詞の韻律の制約からメロディーを計算するシステム (Orpheus)[30] 等がある．深山覚ら [30] の研究で提案された Orpheus では，現在知られている音楽理論に沿ったメロディーが生成されるようなパラメータを使用している．また，Ando Daichi ら [11] の研究ではシステムを使用する人間の好みに沿うような進化がなされる．これらのように，人間の感性を取り入れるようなシステムが結果を残してきた一方で，可能な限り人間の知識を使用しない自動作曲というアプローチも当然考えられる．

近年，機械学習においてディープラーニングという手法が注目されている [1][2][3]．例えば，Quoc ら [3] の研究では，ディープラーニングを用いてネット上の雑多なデータを解析し，ある条件 (例として人の顔に最もよく反応するニューロンを探したところ，そのニューロンを用いたテストデータの解析により 81.7% の精度が得られた．また，そのニューロンをもっとも活性化させる画像を最急降下法で求めるという可能性を見出した．これは，ニューラルネットワークが教師データにはないデータを作成できる能力，つまりは創作の能力を持っていることを示唆している．

本稿では，ディープニューラルネットワークの一部である，LSTM(Long Short-Term Memory) を用いて自動作曲を目指す．ニューラルネットワークによる自動作曲では，初期値のみを与え，時間 t までのデータを用いて時間 $t+1$ の状態を帰納的に予測するという手法が用いられている．また，このタスクの場合，フィードフォワードネットワークと比べてリカレントネットワーク (LSTM はリカレントネットワークの一種) は，過去の入力を記憶する能力があるという点において有利である [19]．本研究もこれに従う．この分野では既に先行研究があるが [6][7]，本研究には自然言語処理の手法を応用した手法を用いるという新規性がある．一つの和音を単語に対応させることによって，音楽を単語のシーケンスデータと解釈し，文章と対応させる．

また，これまでの研究では自動作曲によって生成された曲の数値的評価は難しく，代替手段として作曲に用いたニューラルネットワークが教師データの曲をどの程度再現できるかを評価していた [6][7]．また，生成された曲を音楽の専門家に評価させた研究もある [11][30]．本研究ではデータの前処理がなされているため同等の条件で比較ができないため，新たに生成された曲の評価手段を検討する．

本研究で目指す自動作曲システムの生成物が満たすべき要件は二つあり，教師データをある程度踏襲していること，及びその完全な模倣にはなっていないことである．これは，最低限生成物が創作として成立するために必要な条件である．前述の通り，ニューラルネット

図 1.1: Johann Sebastian Bach, quoted from public domain.

ワークは創作の能力を持っているように見えるが，実際にニューラルネットワークが創作の能力を持っているかどうかというのは自明ではないため，これを定量的に確認する必要がある．

まず，前者の教師データをある程度踏襲しているのは，作曲システムの実用性の面から必要である．ディープラーニングを用いた作曲システムは，教師データを入力して生成した曲を出力する．この時教師データと似ている曲が出力されれば，出力させたい曲と似ている曲を教師データとして採用すればよい，という予想が立てられるため実用性を高めると考えられる．また通常，教師データと全く異なる曲が出力されるニューラルネットワークは学習が正しく行われているとは考えにくい．

これまでの音楽を見ても，偉大な音楽家は他の音楽家から影響を受けると共に後世に影響を与えてきた．例えば，Johann Sebastian Bach は Georg Friedrich Handel の影響を受けていたことが知られており，Bach が Handel の曲を書き写した自筆写譜が残されている．また，Bach の作品である平均律クラヴィーア曲集にも Handel からの借用と思われる箇所が見受けられる [20]．一方，Robert Alexander Schumann は J.S.Bach の影響を強く受けており，その影響は Schumann のあらゆる作品に伺うことができる [21]．このように，音楽家が影響を与え合っていたという事実からもこの「教師データをある程度踏襲していること」という基準の正当性を補強することが出来る．また，教師データの完全な模倣にはなっていないことが創作には必要である．自動作曲システムによって生成された曲が，教師データと同じだった場合，そのシステムは創作という面から見ると，何もしていないのと同じである．また実用上も，著作権の侵害となる可能性があるなどの問題が発生する．

一方で，これら二つの「教師データをある程度踏襲している.」「その完全な模倣にはなっていない.」という要件は相反するものであり，両方を完全に満たすことは出来ない．そのため，図 1.2 に示すような類似度という指標を考えると，どちらかに寄るということは前述

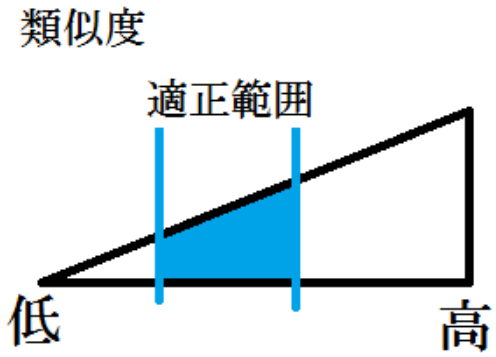

図 1.2: Degree of similarity

の二条件に反する．そのため，中庸であるものが好ましいと考えられる．意図しない著作権侵害問題に対し，中野倫靖ら [26] の研究などでも類似度によるアプローチが行われている．本研究でも類似度を計算するというアプローチを取る．また，音楽が似ているかどうかは一般的には人間の感覚に依る．よってこの類似度は人間の感覚とある程度一致する必要があるため，人間の感覚と比較することによってその妥当性を検証する．

以上から，本研究の目的を二点に纏める．一つはニューラルネットワークを用いた自動作曲をすること，もう一つはこの自動作曲システムの生成物が創作物としての適切であるかどうかの確認である．本研究では生成された曲が音楽的に優れているかの評価はしない．

1.2 本論文の構成

本論文の主な構成を以下に示す．2 章で本研究の主に用いる手法であるディープラーニングについてまとめる．3 章にてニューラルネットワークを用いた自動について関連のある先行研究について触れる．4 章では本研究の提案手法と結果についてまとめ，5 章で結果に対して考察を行う．最後に 6 章に結論を纏める．

第 2章 ニューラルネットワーク

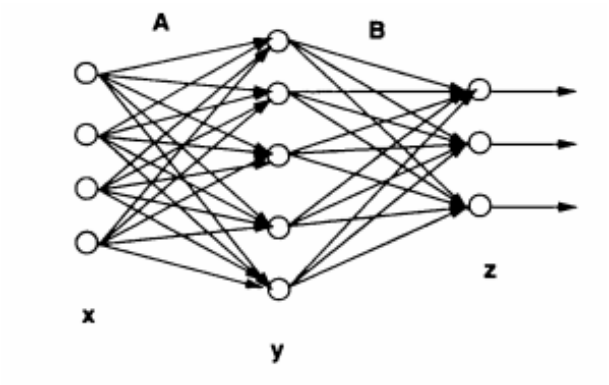

図 2.1: A sample of multi layer perceptron, quoted from [4]figure 1

2.1 多層パーセプトロン

ニューラルネットワークは，他のニューロンから入力を受け取り，出力を行うニューロンの集合体である．一般的には，教師データを入力データとして，望む出力と実際の出力の差を漸近的に縮めていく学習方法を取る．

基本的なニューラルネットワークとして，多層パーセプトロンが挙げられる (図 1)．多層パーセプトロンでは，すべてのニューロンはある層に属し，前の層からの出力を受け取り，次の層へのみ出力する．多層パーセプトロンでは入力信号 $\boldsymbol{x}$ に対し，線形で重み付けをし，隠れ層 $\boldsymbol{y}$ に出力する．また，非線形の表現を行うため，$\boldsymbol{y}$ への出力を非線形関数により処理する．すなわち，隠れ層 $\boldsymbol{y}$ の i 番目のニューロンの入力を y_i とし，入力層のニューロン x_k から y_i への重みを b_{ki} と表すことにすると，

$$y_i = f(\sum_k b_{ki}x_k + b_{0i})$$

(但し，b_0i は bias を示し，定数項の役割を果たす．)

と表せる．出力層 z においても，隠れ層 y を入力として同様の計算が行われる [4]．関数 f の一例としてはシグモイド関数 ($\frac{1}{1+\exp e^{-x}}$) などが用いられる．これを順伝播と呼ぶ．多層パーセプトロンは逆誤差伝搬法 (Back Propagation) という手法により，非線形の問題を解くことが出来る [4][5]． 理想とする出力層のパラメータと上の過程により求めた出力層のパラメータの二乗誤差のサンプル平均を J_{emp} とし，学習率 η をもちいて，重み b を以下のように修正する [5].

$$\boldsymbol{b}_{new} = \boldsymbol{b} - \eta\frac{\partial J_{emp}}{\partial \theta}$$

順伝播と逆誤差伝搬法を繰り返すことによって，学習誤差を単調減少させることが出来る．一方で，それによってテストデータを解析した時の誤差 (汎化誤差) が減少するとは限らない．学習を繰り返すとある時点で汎化誤差が最小値を取り，以降は増加していくことが知ら

れている．これを過学習と呼ぶ．[5] ここで述べたような多層パーセプトロンの問題を解決し，実用性を向上させたニューラルネットワークが一般的にディープニューラルネットワークと呼ばれる．例として，過学習に陥りにくいような良い初期値を探す Auto Encoder や Restricted Boltzmann Machine(RBM)，位置シフトに対して強いネットワーク構造を持つ Convolutional Neural Network(CNN)，時系列データの処理を可能にした Recurrent neural network(RNN) 等の手法が挙げられる．これらを紹介していく．

2.2 Autoencoder

Autoencoder は，出力 $\hat{\boldsymbol{x}}$ を入力 $\boldsymbol{x}$ と同じにする重みを学習する手法である．(図 2.2) この目的を達成するためには入力層 L_1 と出力層 L_3 の要素数は同じとなるが，隠れ層 L_2 の要素数は一般に入力層のものより少なくする (但し，隠れ層の要素数の方が入力層より大きかったとしても，隠れ層がスパースとなるようにペナルティ項を含めることによって学習が出来ることが分かっている．)[22].

学習が終わったのち，出力層を破棄し，入力層から隠れ層への重みを固定する．そして，現在の隠れ層から次の層への学習を同じ手法で行うことを繰り返すことによって，全ての層に対して過学習に陥りにくい初期値を獲得できる．(貪欲法)

このタスクはランダムで独立な入力に対しては非常に難しいが，実際のデータはすべてが独立ではなく，何らかの相関を持っている．少ない次元数で元のデータを表現できるような要素，特徴を学習するのが Autoencoder の目的の一つである．従来のニューラルネットワークの学習では教師データが必要があったが，Autoencoder は教師データを必要とせず，生データを分類することができる．また，学習によって得られた隠れ層を可視化することが出来る．実用的なネットワークに関してはこの手法では難しいが，単純な二層間のネットワークに置いては決定論的に求めることができ，入力層のニューロン x_k から対象の隠れ層のニューロン y_i への重みを b_{ki} と表す．

$$x_k = \frac{b_{ki}}{\sqrt{\sum_k (b_{ki})^2}}$$

上の式を満たす $\boldsymbol{x}$ が隠れ層のニューロン y_i を最大化するので，画像として表現することが可能である [22].

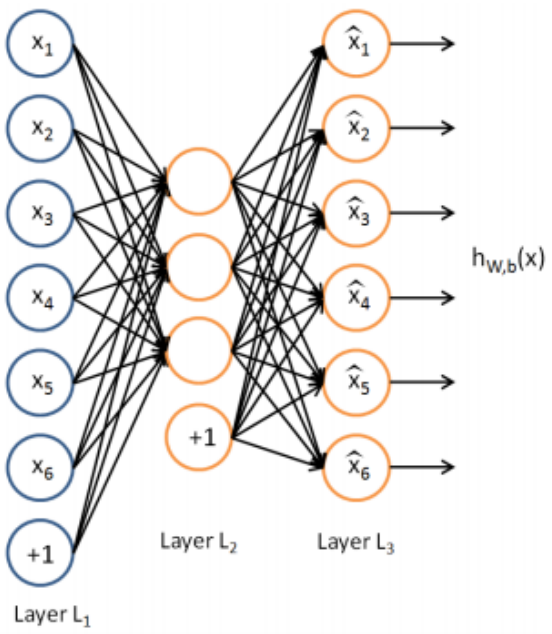

図 2.2: Auto Encoder quoted from [22]13 page

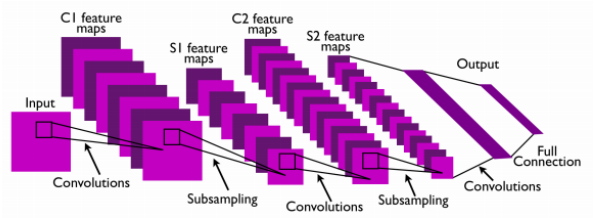

図 2.3: Typical CNN, quoted from [17]Figure 1

2.3 CNN(Convolutional Neural Network)

CNN は，汎化した認識能力を獲得するため，畳み込み層と pooling 層を交互に重ねたネットワークである．基本的な構造を図 2.3 に示す．

2.3.1 畳み込み層

Fukushima ら [16] によって考案された Neocognitron というモデルが元になっている．このモデルでは隠れ層のニューロンが，入力層の一部のニューロン (近隣のニューロン) としか繋がらない．また，Le Cun ら [17] の研究はこのモデルの畳み込み層に RBM による学習を取り入れた．同様に，Quoc ら [3] の研究では Auto Encoder による学習を行っている．この手法では，入力ベクトルをいくつかのサイズの同じ領域に分け (オーバーラップを含む)，それらを全て同じネットワークで解析したデータを出力する．

2.3.2 pooling 層

pooling とは，前述の同種のニューロンを束ねたもの (ニューロンプール) を入力とし，何らかの出力を返す手法を示す．前項の畳み込み層によって得られた局所的な情報を統合し，多層パーセプトロンで学習するのが難しい，位置に対するロバスト性を獲得することが出来る．例として，Max Pooling というものがある．具体例としてサイズ 100*100 のうちに 10*10 のリンゴを含んだ画像を，近隣の 10*10 のリンゴの形に反応するニューロンプールへと入力する．すると，リンゴがある箇所に対応するニューロンが最も大きく反応する．Max Pooling という手法では，このニューロンプールのニューロンのうち最も大きな出力をそのまま出力として採用する．つまり，例に挙げた画像のうち，どこにリンゴがあっても同じ出力が得られると期待できる．

2.4 DBN(Deep Belief Network)

ここでは，DBN を Yoshua ら [2] の研究に沿って紹介する．DBN は，RBM(Restricted Boltzmman Machine の略．Auto Encoder と同じ教師なし学習を担うが，手法が異なる．) の層を重ねたものである．RBM では，ニューロンの活性化はロジスティック回帰によって定められている．ニューロンの状態は 0 か 1 かの離散値で表され，あるネットワークの要素 h_i が入力ベクトルが $\boldsymbol{v}$ の時 1 となる確率 P，入力要素 v_j が出力ベクトルが $\boldsymbol{h}$ の時 1 となる確率 Q を以下のようにおく．但し，b_i, c_j はバイアスを示す.

$$P(h_i|\boldsymbol{v}) = \frac{1}{1 + exp(-b_i - \sum_j W_{ji} v_j)}$$

$$Q(v_i|\boldsymbol{h}) = \frac{1}{1 + exp(-c_j - \sum_i W_{ij} h_i)}$$

ここで，重み行列を W とするとネットワークのエネルギー E は

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{h}' W \boldsymbol{v} - b' \boldsymbol{v} - c' \boldsymbol{h}$$

によって与えられる．(b 及び c はバイアスベクトルであり，学習すべきパラメータ W,b,c を $\theta = (W, b, c)$ と書くこととする．) この E を用いて同時分布 $P(\boldsymbol{v}, \boldsymbol{h})$ は正規化のための係数 Z を用いて

$$P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{v}, \boldsymbol{h})}$$

となる．RBM の対数尤度を最大化するため，偏微分を求める．ここではギブス・マルコフ連鎖を用いる．この $P(\boldsymbol{v}, \boldsymbol{h})$ を $\boldsymbol{h}$ について周辺化することによって，$P(\boldsymbol{v})$ を求め，最尤推定を行う．そのためには $\log P(\boldsymbol{v})$ を θ について偏微分すればよい．連鎖の中の, $\boldsymbol{v}$ の k 番目のサンプルを $\boldsymbol{v}_k$ と置くことにすると，

$$\log P(\boldsymbol{v}_0) = \log \sum_{\boldsymbol{h}} P(\boldsymbol{v}_0, \boldsymbol{h})$$

$$= \log \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}_0, \boldsymbol{h})} - \log \sum_{\boldsymbol{v}, \boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h})}$$

$$\frac{\partial \log P(\boldsymbol{v}_0)}{\partial \theta} = -\sum_{h_0} Q(\boldsymbol{h}_0|\boldsymbol{v}_0) \frac{\partial E(\boldsymbol{v}_0, \boldsymbol{h}_0)}{\partial \theta}$$

$$+ \sum_{\boldsymbol{v}_k, \boldsymbol{h}_k} Q(\boldsymbol{h}_k|\boldsymbol{v}_k) \frac{\partial E(\boldsymbol{v}_k, \boldsymbol{h}_k)}{\partial \theta}$$

となる．(ここでは十分に時間が経ち，真の分布に従うことを仮定するために $k \to \infty$ とする．) この式の第二項に関しては $k \to \infty$ を計算するのは不可能であるが，Gibbs Sampling 等により，十分な数のサンプルを取って平均することで代用する．これは，分布 $P(\boldsymbol{v}|\boldsymbol{h})$ 及

び $Q(\boldsymbol{h}|\boldsymbol{v})$ が既知であることから，適当な $\boldsymbol{v}_0$ に対して P を用いて $\boldsymbol{h}_0$ が計算でき，そこから Q を用いて $\boldsymbol{v}_1$ を計算出来ることを用いてサンプリングを行う手法である．
この偏微分方程式を利用してパラメータを更新していくことで一層についての学習が完了するが，他の層に関しては貪欲法を用いて計算する．即ち，現在の出力ベクトルを入力ベクトルとして利用し，次の出力ベクトルと組み合わせて同じ手順を取る．

ここまでの DBN は離散値を扱ったが，これを連続値に拡張することが可能である．ある値 y とそれに接続しているニューロンのベクトル $\boldsymbol{z}$ を用いて確率分布 $p(y|\boldsymbol{z})$ を考えると，

$$p(y|\boldsymbol{z}) = \frac{e^{ya(\boldsymbol{z})}}{\int_v e^{va(\boldsymbol{z})}dv}$$

となる．(ただし，$a(\boldsymbol{z}) = b + w'\boldsymbol{z}$) 簡単のため，$y$ が $[0,1]$ の区間に含まれるとすると，上式右辺の分母の積分は $\frac{e^{-a(z)}-1}{a(z)}$ となり， y の期待値 $E(y|\boldsymbol{z})$ は

$$E(y|\boldsymbol{z}) = \frac{1}{1-e^{-a(z)}} - \frac{1}{a(z)}$$

これを用いて離散値と同様の手法を行えば連続値を扱うことができる．一方で，明らかに値が 0〜1 の区間内で表せる画像のピクセルデータなどは，離散値の時の確率 P をそのまま連続値として扱ってもよく機能する [23][24].
これらのアルゴリズムを用いて行った実験では，MNIST(0〜9 までの数字の手書きデータ) の分類で，教師あり Deep Net よりも優秀な結果を残している [23].

2.5 Recurrent neural network(RNN)

2.5.1 RNN と通常のニューラルネットワークとの差異

ここでは，Recurrent neural network の基礎的事項を述べ，その発展的な手法であり今回の研究で用いる Long Short-Term Memory という手法について説明する．RNN は，図 2.4 右側に示されるように，フィードバック構造を持つネットワークである．通常用いられる多層パーセプトロンなどのフィードフォワードネットワーク (図 2.4 左側) と対になる．[6][8] 主に時系列データを処理・分類するのに用いられている．例として，[6] 図 2.5 に示すように，入力波形から出力波形を推測するような学習を行うことができる．まず，図 4.A のように与えられた入力データから得られる出力データを教師データに近づける．その後，図 4.B のようにテストとして与えられた入力データから出力データを推測することができる．

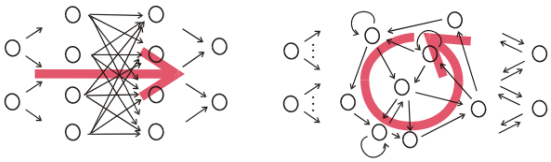

図 2.4: Typical structure of a feedforward network (left) and a recurrent network (right), quoted from [9]figure 1.1

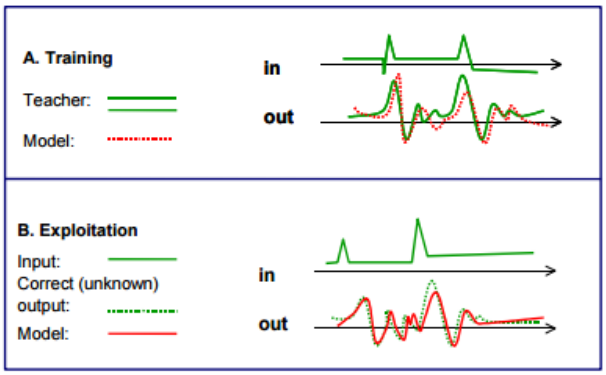

図 2.5: Supervised training scheme, quoted from [9]figure 1.3

2.5.2 RNN の構造

K 個の入力ニューロン，N 個の内部ニューロン，L 個の出力ニューロンを持つニューラルネットワークを考える．時刻 n の，ニューロンの情報を持ったベクトルをそれぞれ $\boldsymbol{u}(n)$(入力ベクトル),$\boldsymbol{x}(n)$(内部ベクトル),$\boldsymbol{y}(n)$(出力ベクトル) とする．この時のコネクションの重み行列をそれぞれ W^{in}(入力層→内部層)，W(内部層→内部層)，W^{out}(※入力・内部層→出力層と原文 [9] にあるが，後述の式と合わせて見ると，入力・内部・外部層すべてを利用している．)，W^{back}(出力層→内部層) とする．これらを用いて，順伝播は以下のような式によって行われる．

$$\boldsymbol{x}(n+1) = f(W^{in}\boldsymbol{u}(n+1) + W\boldsymbol{x}(n) + W^{back}\boldsymbol{y}(n))$$

$$\boldsymbol{y}(n+1) = f^{out}(W^{out}(\boldsymbol{u}(n+1), \boldsymbol{x}(n+1), \boldsymbol{y}(n)))$$

ただし，f 及び f^{out} は活性化関数を示し，sigmoid，tanh，1 などがよく用いられる．フィードフォワードのニューラルネットワークと異なって特徴的なのは，内部層から内部層へのコネクションがあることである．[9]

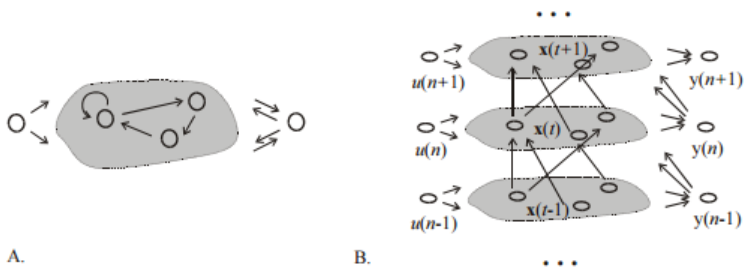

図 2.6: Schema of the basic idea of BPTT. A: the original RNN. B: The feedforward network obtained from it. The case of single-channel input and output is shown., quoted from [9]figure 2.1

2.5.3 基本的な RNN の学習

RNN では主流と言えるような学習方法は確立されていない [9] が，基本的な Back Propagation Through Time, BPTT を紹介する．多層パーセプトロンでの Back Propagation は前に述べた通りであるが，RNN への応用を目指す．まず，通常の Back Propagation はループが無いことを推定しているため，RNN にそのまま応用することはできない．そこで，ループを無くすために，時間方向に RNN を展開するという手法を取る．(図 2.6) これによって内部層→内部層のコネクションは，時刻 n の内部層→時刻 $n+1$ の内部層というコネクションと見なすことができ，ループ構造が解消され，Back Propagation の手法を応用することが可能になる．この手法の問題点は，時間方向に RNN を展開するため，ネットワークの規模が大きくなるため，小規模なネットワーク以外には適用するのは難しいということだ [9]. また，Back Propagation の辿る経路が長くなるため，誤差が指数的に消失，或いは増加する．これは多層パーセプトロンの層の数を増やしたときに起きたのと同じ問題である．このため，タイムラグが 5〜10 単位時間を超えるような性質を学習するのは難しいと言われている [10].

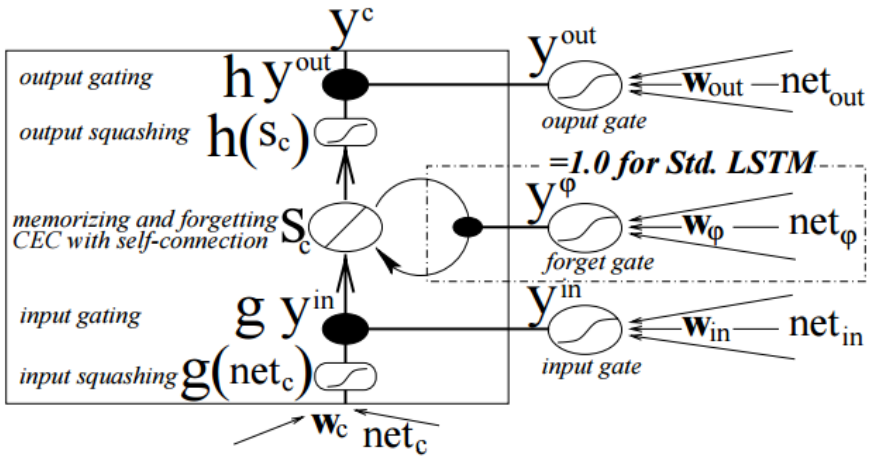

図 2.7: The standard LSTM cell has a linear unit with a recurrent selfconnection with weight 1.0 (CEC). Input and output gates regulate read and write access to the cell whose state is denoted sc. The function g squashes the cell's input; h squashes the cell's output. , quoted from [10]figure 1

2.5.4 Long Short Term Memory

ここでは，Felix A. Gers ら [10] の研究による LSTM を中心に説明する．通常の RNN と違い，1000 単位時間という長周期の性質を学習するのに成功している．図 2.7 のように，通常のニューロンに対し，さらに input gate, forget gate, output gate を追加したユニットを使用する．(元々の LSTM には forget gate は含まれていない．) ここで，説明のため一旦 forget ゲートについては無視する．input ゲートと output ゲートによって CEC(the Constant Error Carousel) という仕組みを作っている．すなわち，ユニットの数値 S_c は

$$S_c(t+1) = S_c(t) + y^{in} g(net_c(t))$$

であり，y^{in} は入力，$g(net_c(t))$ は活性化関数を通した入力ゲートの値である．通常の RNN では誤差の消失が問題となった．ここではユニットにフィードバック機構を付けることで，誤差をそのユニットに留まらせて長期の学習に成功した．ゲートの制御はユニットの値を用いる．即ち，出力と同様に，各ユニットの値の線形結合である．この重みも学習によって更新していく．

この input gate と out put gate を追加した手法は従来の RNN と比べて長期の学習において優れているが，問題点もある．それを解決するために forget ゲートが追加された．問題点として，長期の学習中に，ユニットの数値 S_c が比例的に上昇してしまうことがある，ということが挙げられる．これは，特にユニットに上限下限を設けない場合に，S_c が際限なく上昇して活性化関数の飽和を引き起こす．この事態を防ぐために，forget gate を用い，必要

のない情報を捨てるという選択が出来るようにした．式にすると以下のようになる．

$$S_c(t+1) = y^{\phi} S_c(t) + y^{in} g(net_c(t))$$

この forget gate の状態は，input gate などと同様の手法により決定される．

第 3章 ニューラルネットによる創作の先行研究

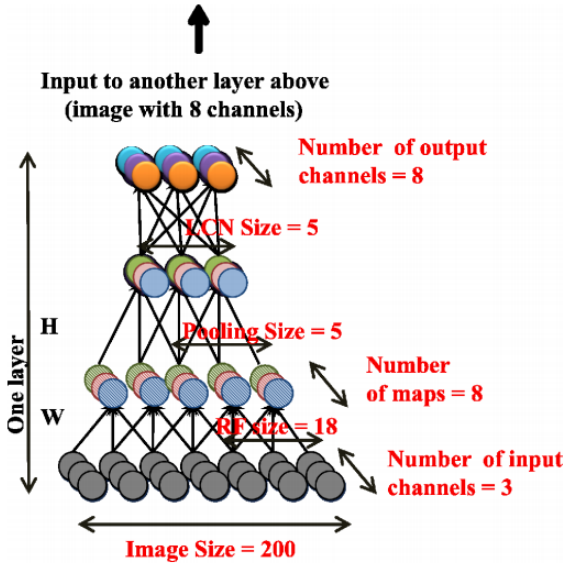

図 3.1: Neural network used in Quoc　2012[3], quoted from [3]Figure 1

3.1　CNN による高レベル特徴の構築

Quoc ら [3] の研究を紹介する．この研究は，ネット上の雑多な画像データからおばあさん細胞 (特定のオブジェクトに強く反応するニューロン) を生成することを目的としている．非常に大規模なニューラルネットワーク (9 層，10 億パラメータ) を利用して，タグ付けされていない画像の特徴を抽出した．ここで用いられたニューラルネットワークを図 2.3 に示す．図 3.1 では，画像データが一層の中で前節で説明した local receptive fields, L2 Pooling, local contrast normalization の三つの工程により処理されている．この出力が次の層の入力となる．この層を 9 層連結して実験で用いるニューラルネットワークとした．(図 3 では画像データが簡単のため一次元データとして表現されているが，実際は二次元である．)

教師なし学習の後，タグ付けされたテストデータを解析し，特定のオブジェクト (論文内では人の顔，猫の顔，人の体について触れられているが，手法は同じためここでは人の顔について述べる) に最も反応するニューロンを探した．その結果，得られたニューロンは 81.7% の精度を示した．

図 3.2 の下側に示すように，得られたおばあさん細胞を最も刺激する画像を最急降下法により解いた．具体的な解くべき問題は以下の式に示す．但し，f はニューラルネットワーク，W, H は学習により獲得したパラメータ，$\boldsymbol{x}$ は入力画像とする．

$$x^* = \arg\max_x f(\boldsymbol{x}; W, H)$$

$$\text{subject to } ||\boldsymbol{x}||_2 = 1$$

図 3.2: Top: Top 48 stimuli of the best neuron from the test set. Bottom: The optimal stimulus according to numerical constraint optimization.　quoted form [3] Figure 3

(注：[3] 原文では arg min となっているが，最大化問題を解くと明記されているため改変した．)

次に，おばあさん細胞のロバスト性を確かめるため，画像に歪みを加えて顔として認識されるかを調べた．その結果，拡大縮小については 0.8 倍から 1.6 倍程度，平面外への回転においては 90 度まで，平行移動については x 軸 y 軸ともに 15 ピクセル程度の移動までは顔と認識され，ある程度のロバスト性があることが示された．また，学習時に与えるデータから OpenCV の顔認識を用いて顔と認識された画像を全て取り除いた所，精度が 72.5% と顔画像を与えた場合に対して減少した．

以上の研究により，教師なし学習によって，人の顔や体を区別するおばあさん細胞が生成されることが確かめられた [3]．この研究は，おばあさん細胞の生成が目的であるが，一方でおばあさん細胞の可視化は創作に近い行為であると考えられる．ニューラルネットワークの持つ創作能力を示唆している．

3.2 LSTM を用いた自動作曲

I-Ting Liu ら [6] の研究である．バッハの曲データを教師データとして用いた学習を行った後，バッハのある曲の最初の音を与えることによって，その音楽を再現するという課題を，LSTM を用いて実行している．

曲データとして，ピアノの音域を基準にした 88 音が，時刻 t に鳴っている，鳴っていないを $1, 0$ で表現したベクトルを使用している．(即ち，時刻 t における $\boldsymbol{x}(t)$ が 88 次元のベクトルである．) このように，次元を減らすなどの前処理を行われていない生のデータを用いたのは，良いネットワークは重みを学習することによってパターン認識を行わなければならないと考えているからであると述べている．このデータ形式を用いて，入力データとして $\boldsymbol{x}(t)$ を与え，$\boldsymbol{x}(t+1)$ を推定するというタスクを行う．ネットワークは結合されている層毎に完全結合であり，[6] input gate, output gate, forget gate の三つのゲートがあると書かれているため，本文には明記されていないが，Gers ら [10] の研究で提案されたネットワークを使用していると思われる．

ネットワークの学習については，RProp という手法を利用している．これは Riedmiller ら [12] の研究で提案された方法であり，back propagation による重みの更新を行う際の，更新量を決める手法である．この手法では，重み w_{ij} の更新量 Δ_{ij} の値を以下のように学習する．

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^+ * \Delta_{ij}^{(t-1)}, if \frac{\delta E}{\delta w_{ij}}^{(t-1)} * \frac{\delta E}{\delta w_{ij}}^{(t)} > 0 \\ \eta^- * \Delta_{ij}^{(t-1)}, if \frac{\delta E}{\delta w_{ij}}^{(t-1)} * \frac{\delta E}{\delta w_{ij}}^{(t)} < 0 \\ \Delta_{ij}^{(t-1)}, else \end{cases}$$

$$where 0 < \eta^- < 1 < \eta^+$$

つまり，時刻 t と $t-1$ の $\frac{\delta E}{\delta w_{ij}}$ の符号が同じなら w_ij の更新量を上げ，違うならば更新量を下げるという手法である．この更新量を用いて，$\frac{\delta E}{\delta w_{ij}}$ と反対の符号に w_ij を変化させるという手法である．[12] この研究では，RProp と BPTT の比較実験を行っている．

データセットとして，J.S. Bach ’ s Chorale midi dataset を用い，その中から教師データとして用いる 4 曲を選んだうえで，一曲ずつ結果が収束するまで学習を行った．後に，ネットワークの能力を調べるため，テストデータの最初の時間のベクトル $\boldsymbol{x}(0)$ を与え，曲全体を再現するというタスクを行った．ここで，曲全体を再現する際にはまず初期値 $\boldsymbol{x}(0)$ を学習が済んだニューラルネットワークに入力し，$\boldsymbol{x}(1)$ を予測する．学習であればこれを実際の $\boldsymbol{x}_{real}(1)$ と比較するが，今回は予測した $\boldsymbol{x}(1)$ に対して閾値 (この論文においては 0.9 とされている．) を設定し，閾値を超えた音について，演奏されていたと判定する．同様に $\boldsymbol{x}(2)$ を予測し，以下帰納的に $\boldsymbol{x}(i)(i$ は任意の自然数) を得ることが出来る．

結果として，F1score にて BPTT による学習を行ったものが 11.84%，RProp によるもの

が 20.29%，Accuracy において，BPTT が 21.03%，RProp によるものが 31.91% という結果を得た．[6] なお，ここで用いられている Accuracy は，用いる音楽データが疎なデータであることを活用し，

$$Accuracy = \frac{TP}{TP + FP + FN}$$

と表す．但し，TP は true positive, FP は false positive, FN は false negative であった音の総数を指す．数として最も多いと考えられる true negative を無視することによってより実態に沿った精度を計算しようとしている．[18]

3.3 RNN-RBM を用いた自動作曲

Nicolas ら [14] の研究では，RBM をリカレントに拡張した The recurrent temporal RBM(RTRBM)，そして RNN-RBM という手法で音楽の再構成を目指している．これは Restricted Boltzmann Machine を拡張した手法である．RBM を用いると，ポリフォニー音楽 (注：さまざまな意味があるが，恐らくここでは同時に二音以上の音が演奏されることのある音楽を指すと考えられる．以降はこの意味で統一して用いる．) の解析に有利であると考えられる．何故なら，ポリフォニー音楽では，ある音が演奏される確率は明らかに同時刻に演奏された他の音に依存しており，通常の RNN はこういった相関を表す能力が低い．一方，RBM ではそのエネルギーを用いた手法により，同時確率の高い組み合わせを選択することができる．I-Ting Liu ら [6] の研究はこれを踏襲しており，行うタスクは同じ形式である．PIANO-MIDI.DE, NOTTINGHAM, MUSEDATA, J.S. Bach’s Chorale midi dataset という四つのデータセットに対し，様々な手法での再構成を行っており，結果を表 3.1 に示す．I-Ting Liu ら [6] の研究と比較可能なデータとして，JSB CHORALES に対する再構成を行った時の Accuracy が，RNN-RBM を用いた時が最大で 33.12% となっている．[14]

Model	Piano-midi.de		Nottingham		MuseData		JSB chorales	
	LL	ACC %	LL	ACC %	LL	ACC %	LL	ACC %
Random	-61.00	3.35	-61.00	4.53	-61.00	3.74	-61.00	4.42
1-Gram (Add-p)	-27.64	4.85	-5.94	22.76	-19.03	6.67	-12.22	16.80
1-Gram (Gaussian)	-10.79	6.04	-5.30	21.31	-10.15	7.87	-7.56	17.41
Note 1-Gram	-11.05	5.80	-10.25	19.87	-11.51	7.72	-11.06	15.25
Note 1-Gram (IID)	-12.90	2.51	-16.24	3.56	-14.06	2.82	-15.93	3.51
GMM	-15.84	5.08	-7.87	22.62	-12.20	7.37	-11.90	15.84
RBM	-10.17	5.63	-5.25	5.81	-9.56	8.19	-7.43	4.47
NADE	-10.28	5.82	-5.48	22.67	-10.06	7.65	-7.19	17.88
Previous + Gaussian	-12.48	25.50	-8.41	55.69	-12.90	25.93	-19.00	18.36
N-Gram (Add-p)	-46.04	7.42	-6.50	63.45	-35.22	10.47	-29.98	24.20
N-Gram (Gaussian)	-12.22	10.01	-3.16	65.97	-10.59	16.15	-9.74	28.79
Note N-Gram	-7.50	26.80	-4.54	62.49	-7.91	26.35	-10.26	20.34
GMM + HMM	-15.30	7.91	-6.17	59.27	-11.17	13.93	-11.89	19.24
(Allan & Williams, 2005)	–	–	–	–	–	–	-9.24	16.32
(Lavrenko & Pickens, 2003)	-9.05	18.37	-5.44	55.34	-9.87	18.39	-8.78	22.93
MLP	-8.13	20.29	-4.38	63.46	-7.94	25.68	-8.70	30.41
RNN	-8.37	19.33	-4.46	62.93	-8.13	23.25	-8.71	28.46
RNN (HF)	-7.66	23.34	-3.89	66.64	-7.19	30.49	-8.58	29.41
RTRBM	-7.36	22.99	-2.62	75.01	-6.35	30.85	-6.35	30.17
RNN-RBM	**-7.09**	**28.92**	**-2.39**	**75.40**	-6.01	**34.02**	-6.27	**33.12**
RNN-NADE	-7.48	20.69	-2.91	64.95	-6.74	24.91	-5.83	32.11
RNN-NADE (HF)	**-7.05**	23.42	**-2.31**	71.50	**-5.60**	32.60	**-5.56**	32.50

表 3.1: Log-likelihood and expected accuracy for various musical models in the symbolic prediction task. The double line separates frame-level models (above) and models with a temporal component (below), quoted from [14]Table 1

第 4章 本研究の提案手法

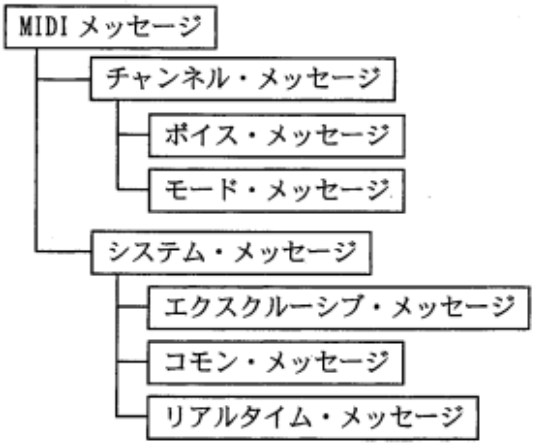

図 4.1: Constitution of midi messages, quoted from [25]figure 3

以上の手法，先行研究を踏まえ，本研究では LSTM を用いて，自動作曲を目指し，その生成物の定量的評価を試みる．本研究の実験は四つの実験からなる．実験 1 では，本研究による提案手法である和音の分割によりデータの次元削減が出来ることを確認する．実験 2 では自動作曲に用いるニューラルネットワークのパラメータを決定する．実験 3 では，自動作曲の生成物の評価を行うための指標について検討する．実験 4 では，実際に自動作曲を行い，実験 3 で得た指標を用いて評価する．

4.1 提案手法及び各手法に対する実験結果

4.1.1 使用するデータ等について

本研究で用いる教師データは MIDI データである．(Nottingham midi dataset)MIDI データをベクトルのデータに変換した後，ニューラルネットワークによる学習を行う．(楽譜に相当するデータ) ここで言う MIDI はデータの規格である．MIDI は図 4.1.1 に示すようなメッセージの集合であり，指示とそれをなすべき時間が示されている．本研究で重要であるのはボイスメッセージである．ボイスメッセージはある音程の音を鳴らし始める，鳴らし終わるという命令を扱う．その他にも MIDI には音色の変更や音量の変更など様々な機能がある [25] が，本研究では用いないため割愛する．MIDI データをベクトルのデータに変換するというのは先行研究でも取られていた手法であり，ピアノの音域を基準にした 88 音が，それぞれ時刻 t に鳴っている，鳴っていないを $1, 0$ で表現したベクトルを使用している (即ち，曲全体を $\boldsymbol{x}$ と表した時，時刻 t における $\boldsymbol{x}(t)$ は 88 次元のベクトルである．)[6]．また，本研究では，MIDI データからベクトルのデータへ変換する際の最小単位を 8 分音符とした．この区切りの中で僅かでもある音が演奏されていた場合，その区切りの中で絶えず演奏されて

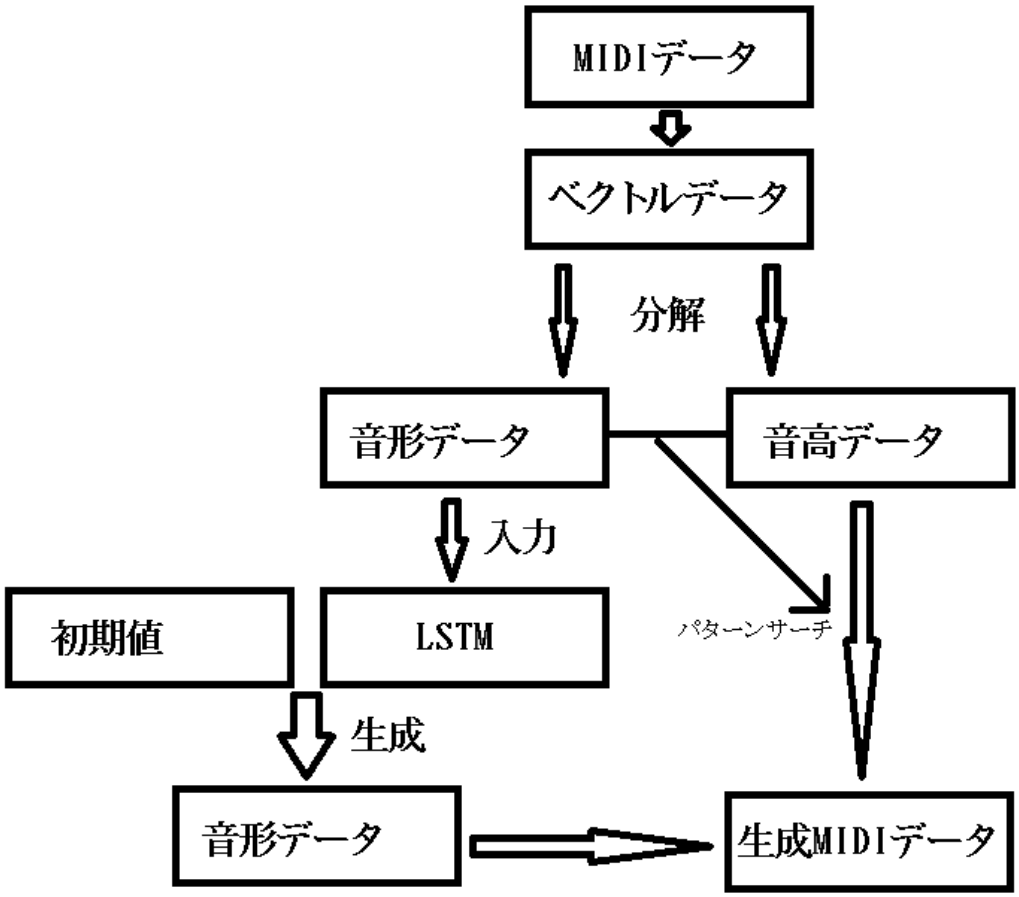

図 4.2: Overview of this auto composing system.

いたと同じ扱いとなる．ここでは例え音量を指定する命令等があったとしても，ボイスメッセージ以外の命令は無視する．本研究では，更にデータに前処理を行う．詳細は次項で説明するが，概要を図 4.1.1 に示す.

4.1.2 実験 1:本研究で使用する，和音の，音高と音形への分割の妥当性についての検討

本研究では，ニューラルネットワークにベクトルデータを与える前に前処理をする．音楽は 図 4.1.2 に示すように，時間方向と音高方向の二次元のグラフであるとみなすことが出来る．これを時間方向で切ると，得られるのは和音である．本研究ではこの和音を音高と音形に分割して考える．和音のベクトルは 88 次元であり，このベクトルは疎であることが多い．(人の指の数の関係から，10 音を超えるような和音は殆ど使用されない．) ここで，和音ベ

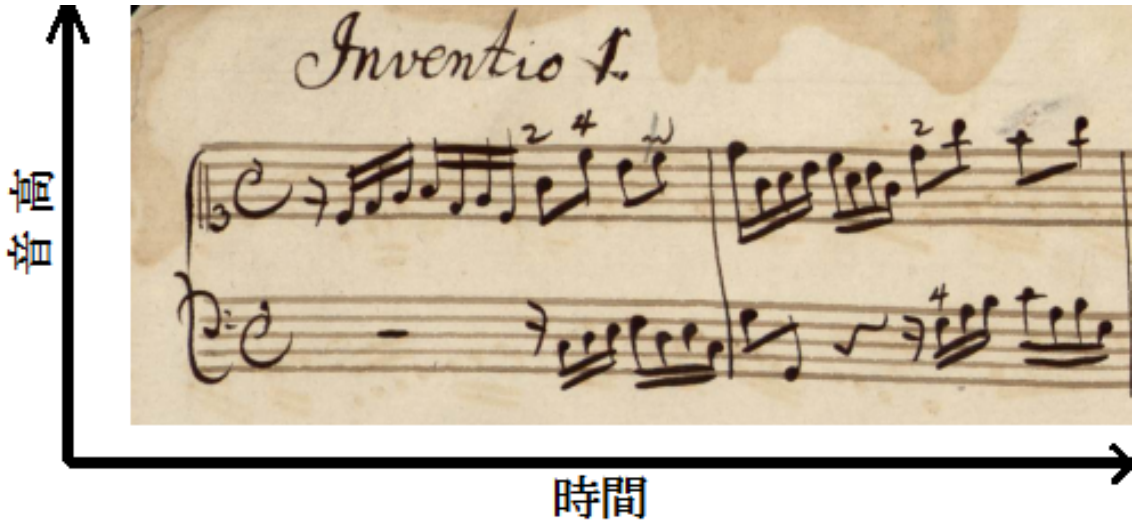

図 4.3: Part of invention 1(J.S.Bach), quoted from public domain

クトルの両端の 0 を省略することを考える．図 4.1.2 のように，基本的な和音であるドミソはベクトル表記だと以下のように表せる．

$$0 \cdots 0100010010 \cdots 0$$

この両端の 0 を無視したものが音形となる．以下に示す．

$$10001001$$

また，左側から 0 を何個省略したかを記録しておくと，元の和音が再現できる．これを音高とする．

この手法を取る妥当性を検討する．和音は最大 88 音の組み合わせであり，非常にパターンが多い．和音であれば 2^{88} 通り，音形であれば，$2^{87}+1$ 通りの音形が存在する．[付録参照] ところが実際に使用される音形は少なく，695 曲の教師データ (Nottingham midi dataset) で使用された音形は 284 通りでしかない．この手法によってデータの次元が大幅に削減出来ていることが分かる．

4.1.3 実験 2:本研究で使用する LSTM のパラメータ選定

本研究では LSTM を用いて自動作曲を行う．ここでは，Jonathan Raiman の theano_ lstm というライブラリを用いた．(これは自然言語処理を目的とした LSTM である．) より良いパラメータを選定するために，実際に行う実験よりも簡単な実験により予備実験を行い，良い精度のものについて実際の実験を行った．ここで言う精度とは，時刻 t までのデータを入力した LSTM により予想した時刻 $t+1$ のデータ $\boldsymbol{x}(t)_{predict}$ が，どの程度実際のデータ

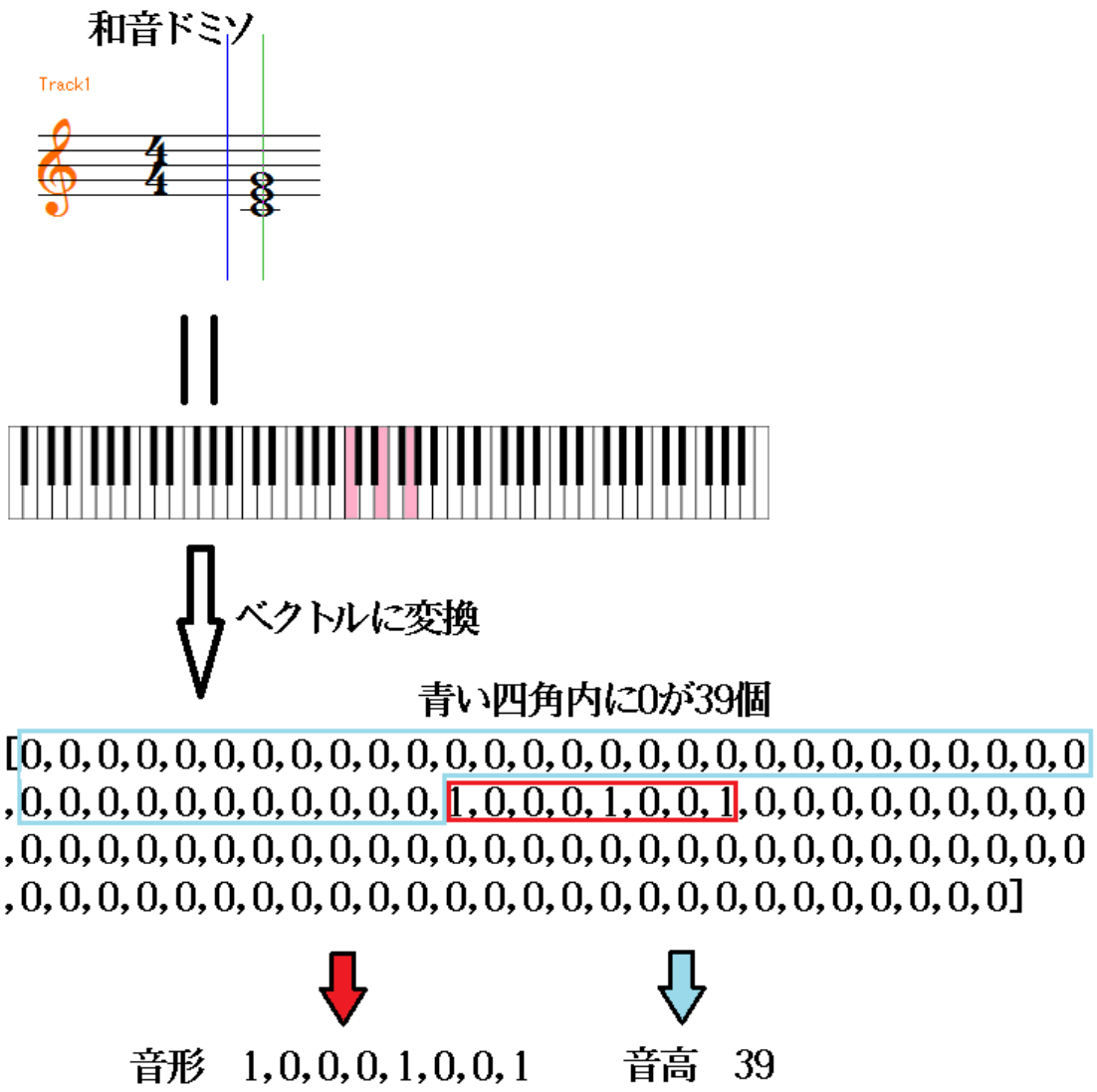

図 4.4: The way to decompose chords into height and shape.

$\boldsymbol{x}(t)_{actual}$ と一致しているかを示す．ここでは，以下のような値を使用している．ここで，予想した

$$Error = \Sigma||\boldsymbol{x}(t)_{predict} - \boldsymbol{x}(t)_{actual}||$$

即ち，各和音 $\boldsymbol{x(t)}$ について，予想と実際の値の差異ベクトルを得て，その距離を誤差とみなす．これを全ての曲について足し合わせたものが Error である．この Error を最小化するような LSTM のパラメータを探す．この Error を，データセットよりランダムに 5 曲選び，5000 epochs の計算をするという条件の元に計算した．また，選定するパラメータとして，Input units の数, Hidden units の数, Memory cells の数の三つがある．図 4.1.3 から，Error の小さいパラメータを選び，100 曲について，10000 epochs の計算を行った．選定したのは，”input units=10, hidden units=50, memory cells=3”, ”input units=20, hidden units=50, memory cells=4”, ”input units=30, hidden units=40, memory cells=2”, ”input units=30, hidden units=40, memory cells=4”, ”input units=50, hidden units=30, memory cells=2”の五つである．

これら図 4.6, 図 4.7, 図 4.8, 図 4.9, 図 4.10 に示す結果から，最も 10000Epochs の計算が終了した後の Error が低いパラメータを選んだ．以上からパラメータは input units=20, hidden units=50, memory cells=4 と決定した．

Input Units	Hidden Units	Memory Cells	Error
10	10	1	624.2491
10	10	2	346.8089
10	10	3	154.3135
10	10	4	860.9517
10	20	1	67.6858
10	20	2	14.30367
10	20	3	31.80495
10	20	4	24.19087
10	30	1	61.85794
10	30	2	10.64114
10	30	3	14.58766
10	30	4	10.30479
10	40	1	29.74202
10	40	2	10.62649
10	40	3	15.72361
10	40	4	6.128946
10	50	1	11.60959
10	50	2	6.297086
10	50	3	4.816673
10	50	4	6.232917
20	10	1	508.3974
20	10	2	362.9688
20	10	3	187.1138
20	10	4	198.2157
20	20	1	50.57842
20	20	2	17.87418
20	20	3	13.93409
20	20	4	14.07486
20	30	1	14.54213
20	30	2	9.177566
20	30	3	24.84612
20	30	4	8.941128
20	40	1	13.05518
20	40	2	12.07904
20	40	3	10.45652
20	40	4	10.38917
20	50	1	7.749445
20	50	2	18.37099
20	50	3	10.53419
20	50	4	9.111484
30	10	1	561.7714
30	10	2	254.1069
30	10	3	53.61965
30	10	4	134.2307
30	20	1	34.2354
30	20	2	11.68768
30	20	3	10.81199
30	20	4	11.42491
30	30	1	11.64348
30	30	2	11.00691
30	30	3	8.955603
30	30	4	10.32983
30	40	1	17.14379
30	40	2	4.765078
30	40	3	9.010513
30	40	4	4.779167
30	50	1	12.00705
30	50	2	7.625987

図 4.5: Error calculated with various parameters.

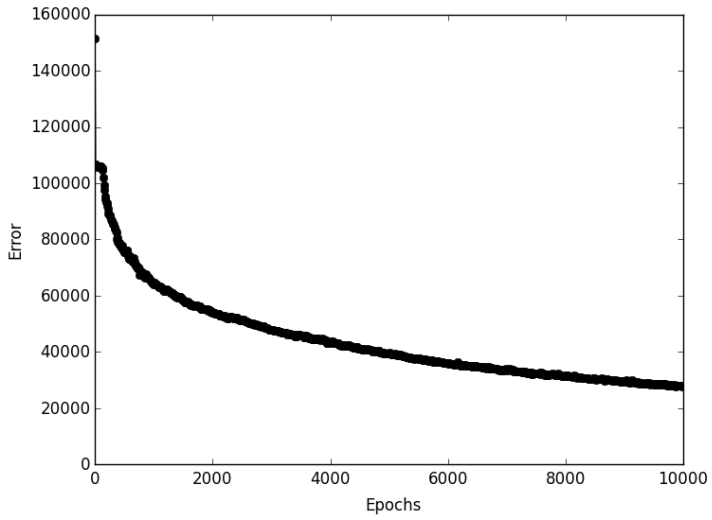

図 4.6: Error calculated with ”input units=10, hidden units=50, memory cells=3”.

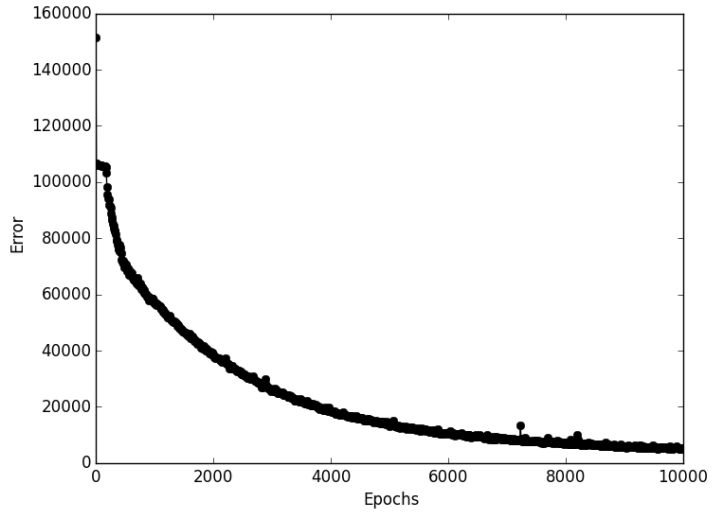

図 4.7: Error calculated with ”input units=20, hidden units=50, memory cells=4”.

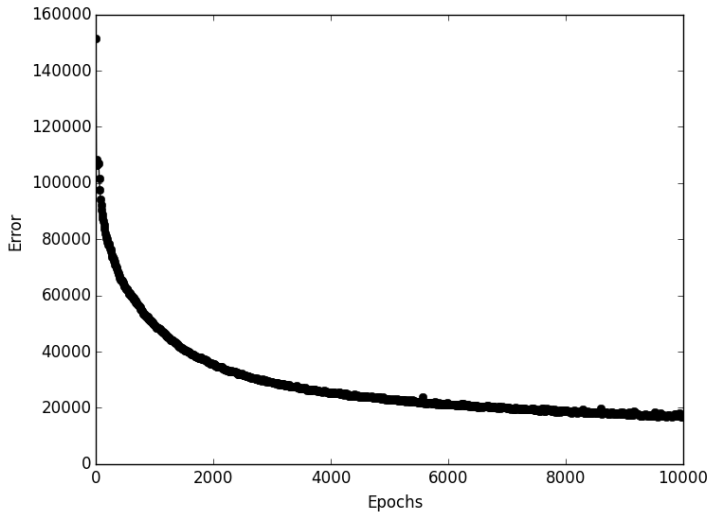

図 4.8: Error calculated with ”input units=30, hidden units=40, memory cells=2”.

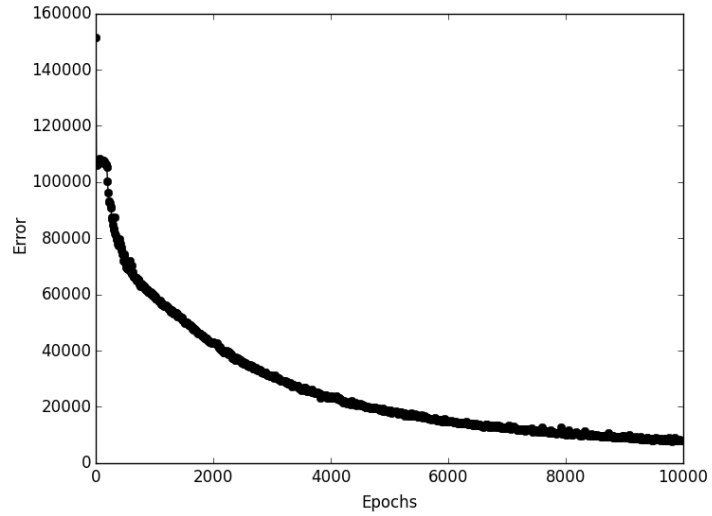

図 4.9: Error calculated with ”input units=30, hidden units=40, memory cells=4”.

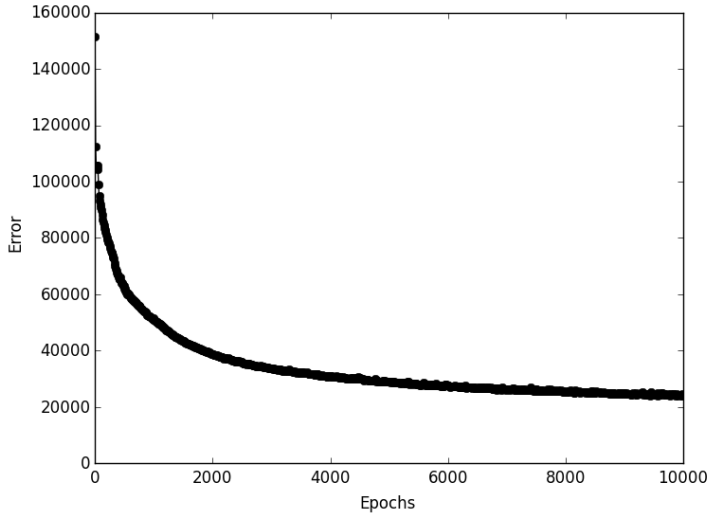

図 4.10: Error calculated with ”input units=50, hidden units=30, memory cells=2”.

4.1.4 実験 3:ニューラルネットワークが創作能力を持つかどうかについての検証

自動作曲という分野において，それを学術的に評価するのは困難を伴う．また，本研究では，ニューラルネットワークにおける自動作曲というテーマを扱っているが，ニューラルネットワークが創作する能力を持っているかどうかについては必ずしも自明ではない．先行研究では，音楽の再現というタスクを実行することによって，元の音楽と比較することによって精度の定量的な評価をする手法 [6]，音楽の専門家に評価させる手法 [11] 等が行われている．本研究では，研究の目的である教師データをある程度踏襲していること，及びその完全な模倣にはなっていないことという二点を満たしているかどうかについて，類似度という概念を導入して定量的な評価を試みる．何故なら，教師データを踏襲する，完全な模倣にならない，という二点の目標は相反している．そのため，両方を完全に満たすことは出来ないため，数値的な指標の存在が仮定できる．この類似度において，本研究では中庸であることを目指すとともに，どの程度の範囲が望ましい範囲であるのかを人間の感覚と比較して検証する．ここまでで，創作の必要条件として類似度という指標が存在し，それが中庸であることが必要である，という仮説を立てた．ここで，二つの類似度 R,D を考える．曲 u の時刻 t における音形を $y_u(t)$ とおき，曲 u と曲 v の内短い方の長さを N とおく．

$$R = \max\{\frac{1}{N}\Sigma_{t=0}^{N}\delta_{\boldsymbol{y_u(t)},\boldsymbol{y_v(t+i)}}|i\}$$

但し，δ はクロネッカーのデルタである．この式は，曲 u と曲 v の時間シフトを無視した時に，どの程度共通の音形が用いられているかを示す．最大で 1，最小で 0 の値を取る．更に，

類似度 D では音形だけではなく音高を用いる．曲 u の時刻 t における音高を $z_u(t)$ とおき，

$$D = \max\{\frac{1}{2N}(\Sigma_{t=0}^{N}\delta_{\boldsymbol{y_u(t),y_v(t+i)}} + \Sigma_{t=0}^{N}\delta_{\boldsymbol{z_u(t),z_v(t+i)}})|i\}$$

類似度 R は D と比べて音高を使用しないため情報量が少なく，精度も低いと考えられる．しかし，本研究は音形を決定した後に音高を決定するという形を取るために，音形のみで類似度を判断できるのが望ましいためこちらも検討する．これら類似度 R, D は，ペアワイズアライメント [29] を参考に改変したものである．詳細は考察に記す．これらが人間の感覚と一致するかどうかを検証するためにアンケート調査を行った．サンプル数は 10 であり，二曲を一度ずつ聴いた後に，それら二曲の類似度を三段階 (選択肢の上の方から「似ていない，別の曲である」「似ているが別の曲である」「殆ど同じ曲である」) で判定するという形式を取った．それぞれ類似度の異なる，Nottingham midi dataset から抽出した 9 組 18 曲を選定し，本研究で使用した類似度 R と，人間の感覚がどの程度相関があるかを調査した.「似ていない，別の曲である」を 1,「似ているが別の曲である」を 2,「殆ど同じ曲である」を 3 とした．同一の曲セットに対し，アンケートの解答結果の平均を取った上で (これを人間の感覚による類似度とする)，これを今回使用した類似度 R, D によるものと比較した．図 4.11 に示す類似度 R と人間の感覚による類似度との相関係数を計算した所，0.50 という値を得た．類似度 R は，人間の感覚と強い相関があるとは言い切れない．一方で，図 4.12 に示す類似度 D については，人間の感覚による類似度との相関係数が 0.92 という結果になった．サンプル数は少ないが，10 件の解答を，平均を取らずに類似度 D との相関を調べたところ，最低のもので 0.52 であり，図 4.13 に示すように全体として高い相関を示した．以上を以て，類似度 D と人間の感覚による類似度には強い相関があると言える．

4.1.5　実験 4:LSTM による自動作曲

実験 1 の項で述べたように，本研究では和音を音形と音高に分割した．まずはそのうち音形を扱う．LSTM に教師データとして，Nottingham midi dataset の内 40 曲からなる音形データを与え，10000 epochs の計算を行った．以下，音形データを得る方法について説明する．

教師データをそれぞれ時間方向に切ることによって和音データを得る．和音データを音形と音高に分割する．ここで，音高はスカラーのデータであるが，音形はベクトルのデータである上，サイズは可変である．このままでは扱いにくいため，種類毎に番号を振る．この番号の数字に音楽的な意味は無いため，ニューラルネットワークに入力する時にはベクトル (n 次元，但し n は音形の種類の数) に変換する．

例として 5 種類の内の 3 番目の音形は以下のようなベクトルで表される．

$$\{0, 0, 1, 0, 0\}$$

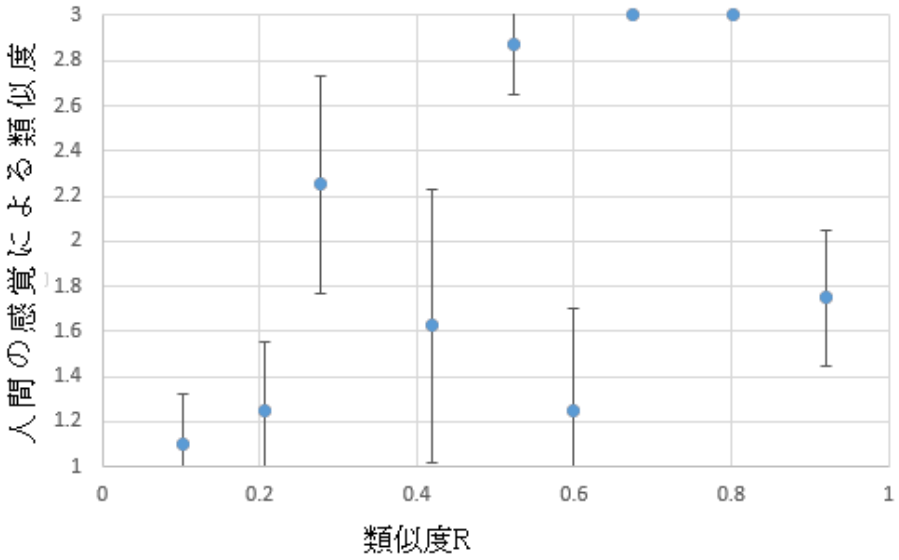

図 4.11: Human feeling and calculated similarity R.

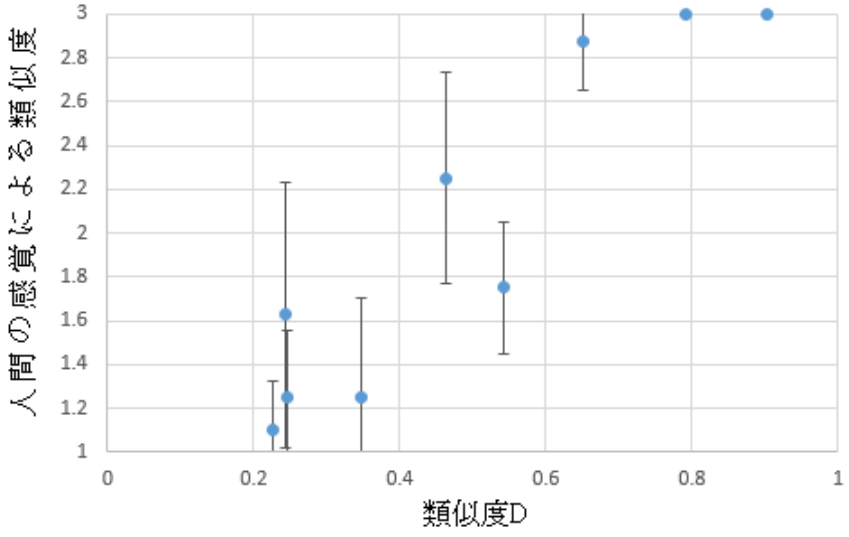

図 4.12: Human feeling and calculated similarity D.

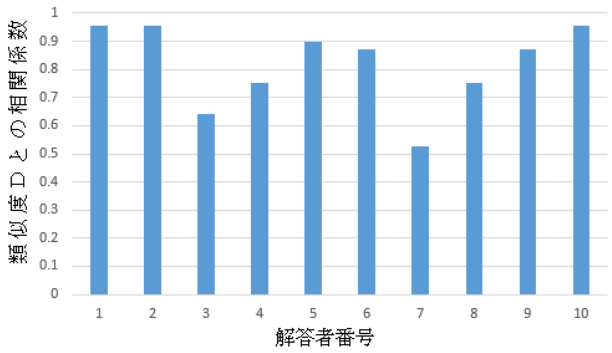

図 4.13: Correlation coefficients between individual data and calculated similarity D.

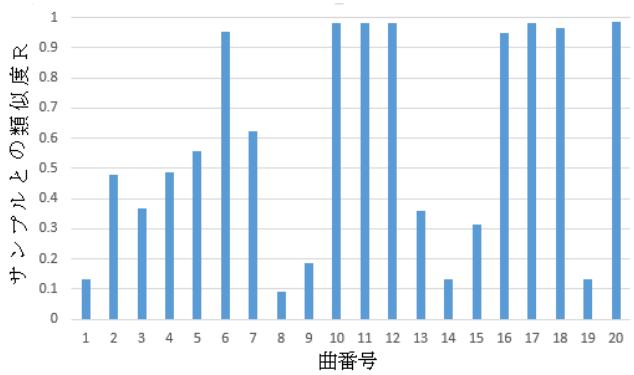

図 4.14: Degree of Similarity R between teacher data and composed data in this experiment.

上例のように，t 番目の音形は，t 番目の成分のみが 1 であり，残りの成分が 0 であるベクトルで示される．

学習によって得たニューラルネットワークに異なる 20 種類の初期値を与え，それぞれに音形の列を得た．ここで音形の列を得る方法は前述した Liu ら [6] の手法にならうが，データの形式が異なる (Liu ら [6] の研究で用いられているデータは鳴っている音を 1, 鳴っていない音を 0 で表した 88 次元のベクトルである．) ため，閾値を設ける代わりに，ベクトルの成分の中で最大の成分を持つ音形を採用する．これらを評価する手法として，実験 3 で示した類似度 R を用いる．生成した 20 曲の音形のみのデータについて，学習に用いた教師データに含まれる曲との類似度 R の内最大のものをを計算した結果を図 4.14 に示す．また，音形と音高のうち，音高についての研究は不十分な状態である．一方で，生成されたデータが音楽として成立するためには音高を決定する手段についての研究が必要である．そのため今回は仮の手段として，音形のみのデータから音高と音形のデータである音楽を作成するために以下のような手法を用いた．

1. 音高のデータは，すべて最初の音との差異という形で相対的な値として扱う．また，音高についてはオクターブの違いを無視する．作成する曲の音高を Z_t と置き，音形を Y_t と置く．
2. 教師データから切り出した音形と音高データを YT_t, ZT_t と置く．これらと，作成する曲の一致率が高い部分を探し，参考にするという手法を取る．式は以下のようになる．一致率を I と置き，

$$I = \frac{1}{2N+1}\{\Sigma_{i=1}^{N}(\delta_{YT_{t-i},Y_{t-i}} + \delta_{Z_{t-i},ZT_{t-i}}) + \delta_{YT_{t-i},Y_{t-i}}\}$$

但し，N は最初 19 とした．
3.2 のステップで，最も一致率の高かった教師データにて $I > 0.9$ であった場合，その教師データが次に演奏する音高 ZT_t を採用する．一致率がそれに満たない場合，N=18,17... と検索する単位時間を減らして 3 のステップをもう一度行う．
4. これを繰り返して全単位時間について音高を決定する．
纏めると，教師データから，作成中の曲と一致度の高い部分を検索し，次の音高を検索によって発見した教師データから採用するという手法である．

例を挙げると，

$$\text{音高} : \{0,0,0,5,5,k\}, \text{音形}\ \{1,1,0,2,2,3\}$$

というデータがあり，k を決定したいとする．但し，音高，音形のベクトルにおいて，それぞれ j 番目の成分は時系列的に一致しているとする．教師データに以下のようなデータがあり，

$$\text{音高} : \{0,0,0,5,5,0\}, \text{音形}\ \{1,1,2,2,2,3\}$$

教師データの中でこの一致率 I が最も高かった場合 (この場合，$I = \frac{10}{11} > 0.9$ であるため，前項 3 の条件を満たしている．)，$k = 0$ であると決定する．

この手法により生成された曲の一部を図 4.15, 図 4.16 に示した．図 4.17 などに示される曲の番号で，図 4.15 は 9 番に，4.16 は 3 番に対応する．また，図 4.14 に示された音形のデータについて，音高を付与して生成された曲の類似度 D を図 4.17 に示した．

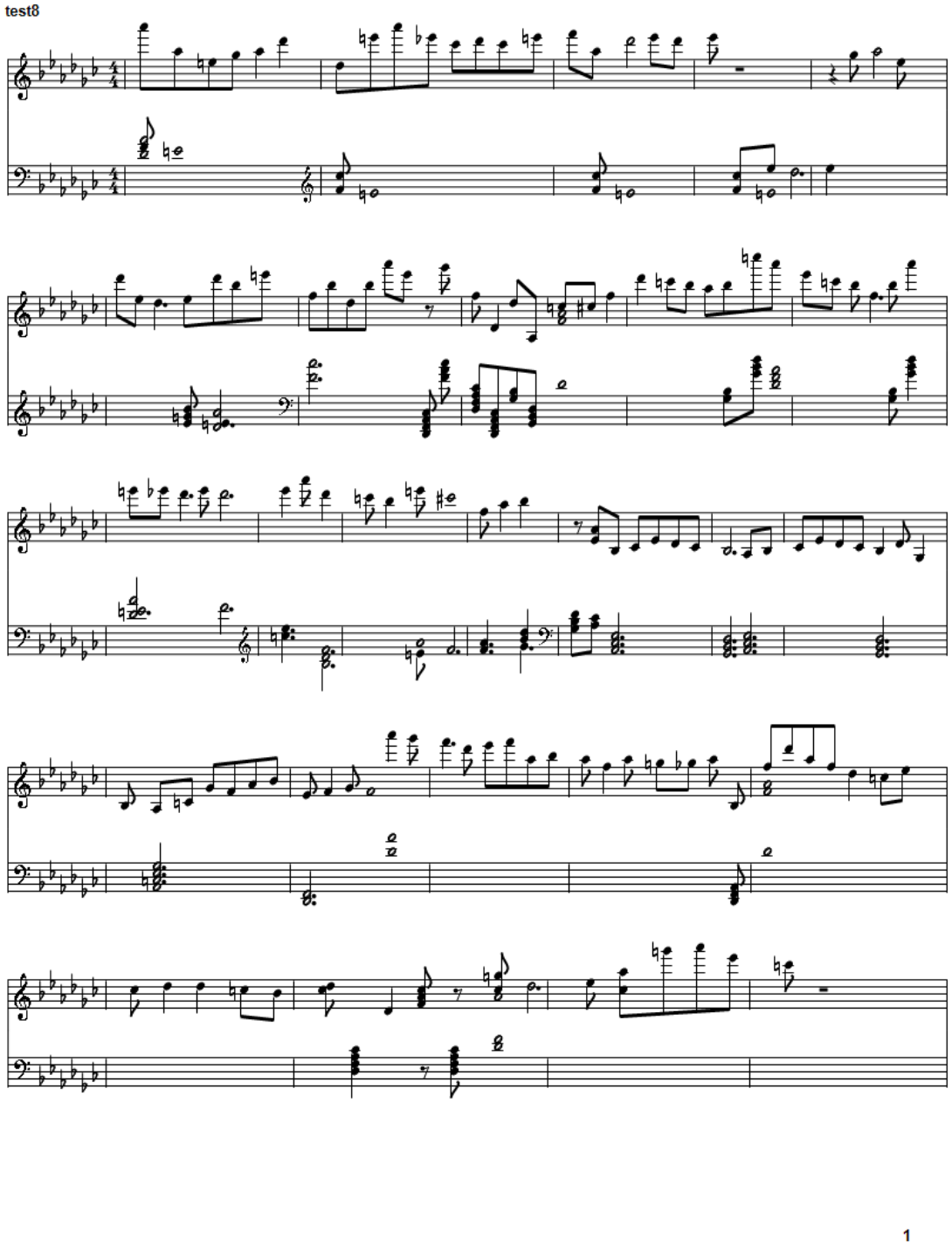

図 4.15: A sample of composed music 1.

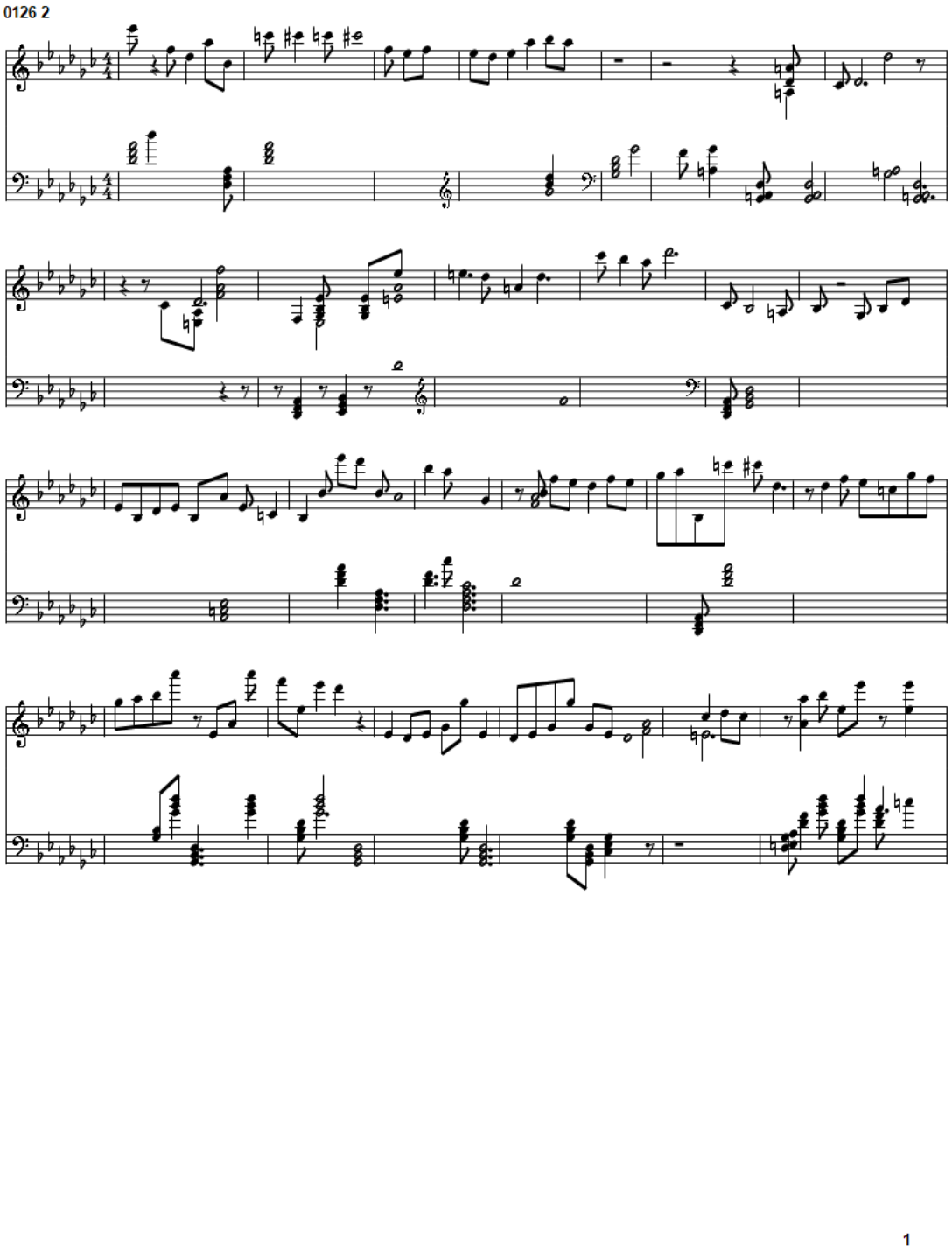

図 4.16: A sample of composed music 2.

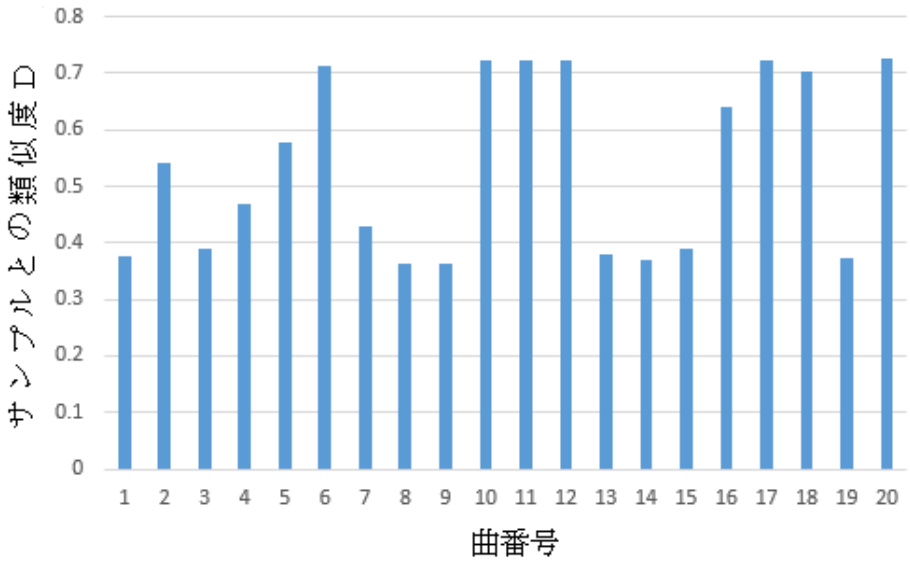

図 4.17: Degree of Similarity D between teacher data and composed data in this experiment.

第 5章 考察

5.1 和音の分割についての考察

実験 1 では，和音を音形と音高に分割するという手法を取った．これによって $2^{87}+1$ 通り存在する音形の内，Nottingham midi dataset の 695 曲の教師データ (Nottingham midi dataset) で使用された音形は 284 通りであった．LSTM による自動作曲の先行研究 [6] では，初期値のみを与え教師データを再現するというタスクに対し，31.91% の精度を得ている．データに前処理を加えていない先行研究と前処理を加えた本研究との比較では条件が異なるが，本研究では，高いもので教師データとの類似度 D で 0.7 を超えており，これは参考にした曲と位置シフト (転調)，時間シフトを無視することによって少なくとも 70% が単一の教師データと一致したということを意味する．和音を音形と音高に分割するという手法により，より良い精度の学習が出来ている．また，Fujishima ら [27] の研究ではコード進行を分析するというタスクに取り組んでいるが，扱ったコードは 27 種類であった．これに単音メロディーのパターンが加わり，コードと単音のメロディーの組み合わせを考えても，高々 $2376(27*88)$ パターンであり，$2^{87}+1$ と比較すると十分小さいと言える．そのため，コードと単音のメロディーの組み合わせというパターンに収まる楽曲であれば，この手法によって良い精度の学習が出来ると期待できる．また，本研究では自然言語処理に用いる LSTM を流用した．これは，和音の分割によって得られた音形と単語の対応関係があるため可能になった手法である．自然言語処理の分野で進歩があった場合，音楽の分野に応用できる可能性がある．

5.2 ニューラルネットワークの創作能力の有無についての考察

実験 3 によって得られた結果から，本研究で検討した類似度 D がある程度人間の感覚と相関があることが示された．

ここで創作に必要な 2 条件 (教師データをある程度踏襲している，完全な模倣にはなっていない) を満たすためには，類似度が一定の範囲にあることが望ましい．

ここで，今回の条件を満たすためにはアンケートに用いた選択肢にして「似ているが別の曲である」(2) と判定されるのが最も望ましい．一方で「殆ど同じ曲である」(3) と「似ていない，別の曲である」(1) を比較した時,「殆ど同じ曲である」(3) と判定されると著作権等の問題が発生する可能性があるため，どちらかと言うと「似ていない，別の曲である」(1) と判定される方が望ましい．以上の結果から，人間の感覚による類似度が 1.5 から 2 の範囲に入ることが望ましいとする．この指標はシステムを利用する目的によって変わりうるが，本研究ではこれを以て LSTM による自動作曲の結果を評価する．

類似度 D を回帰分析した結果が図 5.2 に記載されている．但し，エラーバーは 95% 信頼区間を示す．また，類似度 D は人間の感覚と強い相関 (相関係数 0.92) を持っているため，

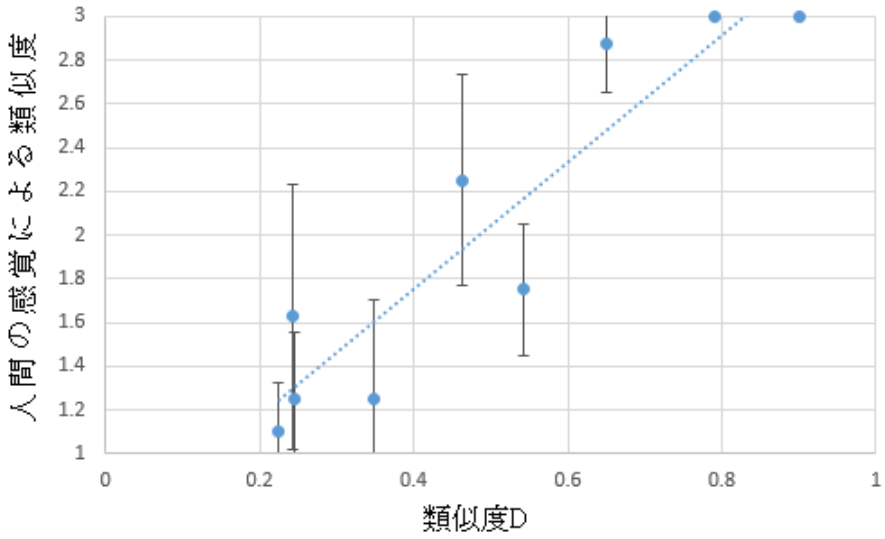

図 5.1: Human feeling and calculated similarity D with regression line.(Error bars means the 95 % confidence interval.)

簡単のため一次の直線で近似した．ここから人間の感覚による類似度 1.5 から 2 に相当する類似度 D は 0.3〜0.5 となる．(元々の仮定に確固たる根拠が無いため，有効数字は 1 桁とした．)

また，こういった配列同士の類似度の計算には動的計画法によるペアワイズアライメントが用いられることがある．ペアワイズアライメントとは，二つの配列の対応関係を求める手法である．それぞれの配列に空白を挿入することが出来，一致している場合 2 点，一致しない場合 −1 点，片方に空白が挿入されている場合 −2 点などのように得点を設定し，得点が最高になる対応関係を探す．

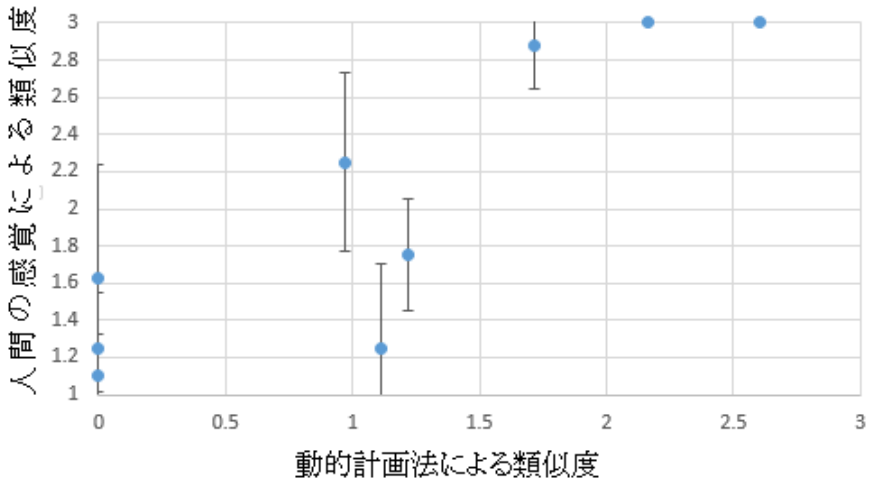

図 5.2: Human feeling and calculated similarity by pair wise alignment.

例えば，

$$GCCAUUG$$

$$GCCUCG$$

という二つの配列があった場合，最適なアライメントは，

$$GCCAUUG$$

$$GCC - UCG$$

(但し，-は空白を示す) であり，得点は 2 点 (一致 5，不一致 1，空白 1) である．最適なアライメントを求める手法として，Smith ら [29] の研究による Smith-Waterman アルゴリズムを本研究では用いた．[29] また，本研究の条件に対応させるため，最高の得点を得るように配列の一部のみを用いることが出来るようにした．(この条件下では，前述の例では，二つの配列から GCC をそれぞれ抜き出し，全て一致するので得点は 3 点となる．) 音高と音形という二種類のデータを比較する必要があるため，両方一致した場合 3 点，片方一致した場合 1 点，両方とも一致しなかった場合 −1 点，片方に空白が挿入されている場合 −2 点という点数に設定した．また，曲の長さに左右されないように，類似度はペアワイズアライメントによって得られた類似度を短い方の曲の長さで割ったものとした．ペアワイズアライメントによって得られた類似度を人間の感覚と比較した結果を図 5.2 に示す．相関係数は 0.87 であった．この手法においても人間の感覚と高い相関があるが，本研究で使用した類似度 D と同程度であり，類似度 D は，このペアワイズアライメントにおけるアライメントの得点を，

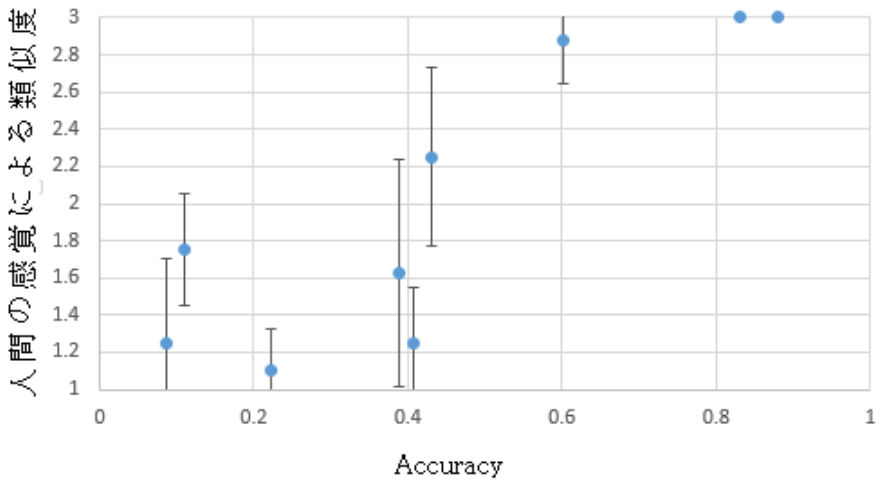

図 5.3: Human feeling and calculated accuracy used in [6][14].

両方一致した場合 1 点，片方一致した場合 0.5 点，両方とも一致しなかった場合 0 点，片方に空白が挿入されている場合 $-n$ 点 (但し，$n \gg 1$) と設定した場合と対応する．

また，本手法では音形と音高への分割というデータの前処理を行ったが，その処理を行う前のベクトルデータ (ピアノの音域を基準にした 88 音が，時刻 t に鳴っている，鳴っていないを 1,0 で表現したベクトル [6] を時系列方向に並べたもの) で類似度が考えられるかについて検討した．先行研究である，Nicolas　2012[14]，Liu　2014[6] で生成された楽曲がどの程度教師データと一致するかを指標 Accuracy として表していたが，これを用いる．(詳細は先行研究の章を参照)

$$Accuracy = \frac{TP}{TP + FP + FN}$$

また，本研究で用いた類似度 D の条件と合わせるため，音高方向のシフト，時間方向のシフトを全パターン計算し，最大のものを採用した．結果を図 5.3 に示す．相関係数は 0.86 であった．即ち，類似度の計算に関しては，音形と音高の分割という手段を取っていない場合であっても，人間の感覚と相関の高い指標を得ることが出来るため，応用が考えられる．将来の課題としては，本研究のアンケートでは，Nottingham midi dataset という限られたデータセットの中での類似度を測ったが，一般の音楽に対して同じように類似度が人間の感覚と相関するのかどうかについて検証する必要がある．

5.3　LSTM による自動作曲についての考察

本研究では LSTM を利用したが，これは自然言語処理の技術を利用できるためである．音形は単語と対比可能であり，音楽は単語の羅列，つまり文章と対比可能である．LSTM を

利用することにより，人間が音楽理論をコンピュータに教えることなく，教師データを与えるだけで自動作曲が可能である．ニューラルネットワークを用いない先行研究 [11][30] では，和声の遷移確率の設定や，好ましい旋律の人間による選択など，システムに人間の感性が取り入れられている．ニューラルネットワークを用いることによって，音楽データから直接人間の感性を取り入れられる可能性がある．

図 4.17 に示す自動作曲システムの生成物の一部の類似度 D はこの範囲に入っている．本研究の自動作曲システムは創作をする能力を持っていると主張できる．

また，この自動作曲システムは，図 4.17 に示すように，これらの要件を満たさない楽曲を多く生成している．

図 5.4, 5.5 はそれぞれ本システムが生成したデータと，それと最も類似度が高かった教師データとなっている．類似度 D は 0.7239 であり，目安となる 0.3〜0.5 の範囲よりも高い．これらの音楽は極めて似通っており，自動作曲の生成物としては不適切である．本手法では音高をパターンサーチする手法を取っており，類似度 D は本来の LSTM の学習能力よりも高くなってしまう可能性があるが，これらの曲の類似度 R(音形のみの類似度) は 0.9801 であった．これは自動作曲の生成物としては不適切であるが，一方で LSTM の学習能力を示している．

生成物は中庸な類似度を持ち，一方でシステムは高い類似度を持つ音楽を生成できる能力を持つべきというのは一見矛盾する主張である．しかし，これを人間に置き換えると，先人の作品を勉強して再現できるようになり，それを踏まえて音楽を作るべきである，というのは自然な主張である．高い類似度を持つ音楽を生成できる能力が必要であるとすると，乱数によって教師データにノイズを加える (この手法によっても中庸な類似度は達成できると考えられる) ような手法よりもニューラルネットワークによる学習の方が望ましい．

類似度の要件を満たさない楽曲を出力することは意図しない著作権の侵害を引き起こすことがあり危険である．しかし類似度の定量化によって，要件を満たさない楽曲の出力を取りやめるという機能の実装が可能になる．これによって自動作曲システムの実用性を引き上げることが出来る．

また，実験 4 では実験 1 によって得た音形間の関係について全く考慮しなかった．音形間の関係を考慮することによって学習の精度は更に向上し，類似度の判定の精度も向上すると考えられる．なお，本研究では教師データの音形データからネットワークを作り，regular equivalence[28] によって構造的に音形の類似度を計算する試みを行い，結果として失敗したことを付記しておく．(教師データの音形を node として，時系列的に前後の音形に path が繋がっているものとしてネットワークを構築し，各 node 間の regular equivalence に有意な結果が得られなかった．)

音高のデータは，すべて最初の音との差異という形で相対的な値として扱った．これは，人間の相対音感の特性を利用した．峯松信明ら [31] の研究で述べられているように，ソミ

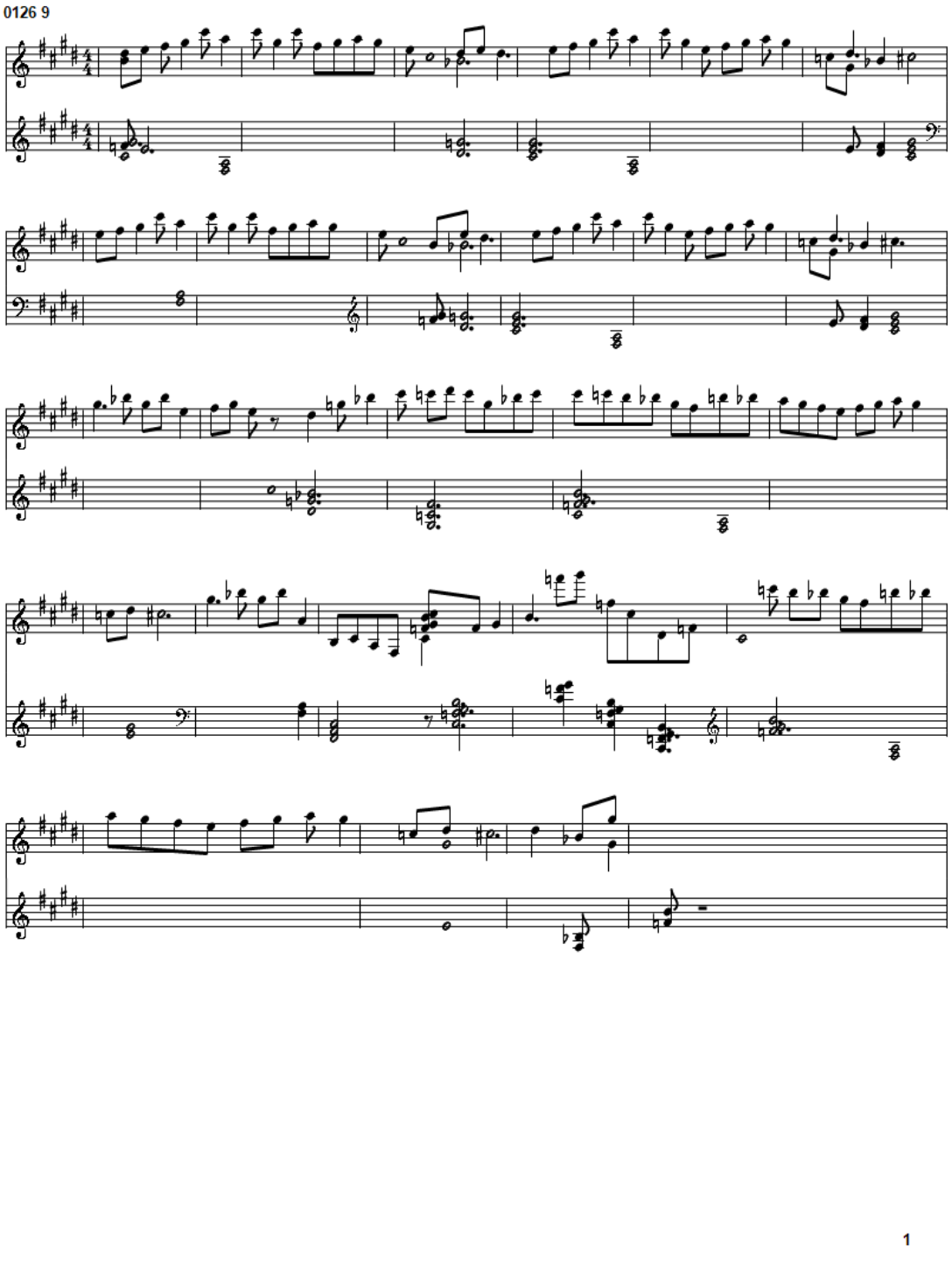

図 5.4: A sample of composed music 3.

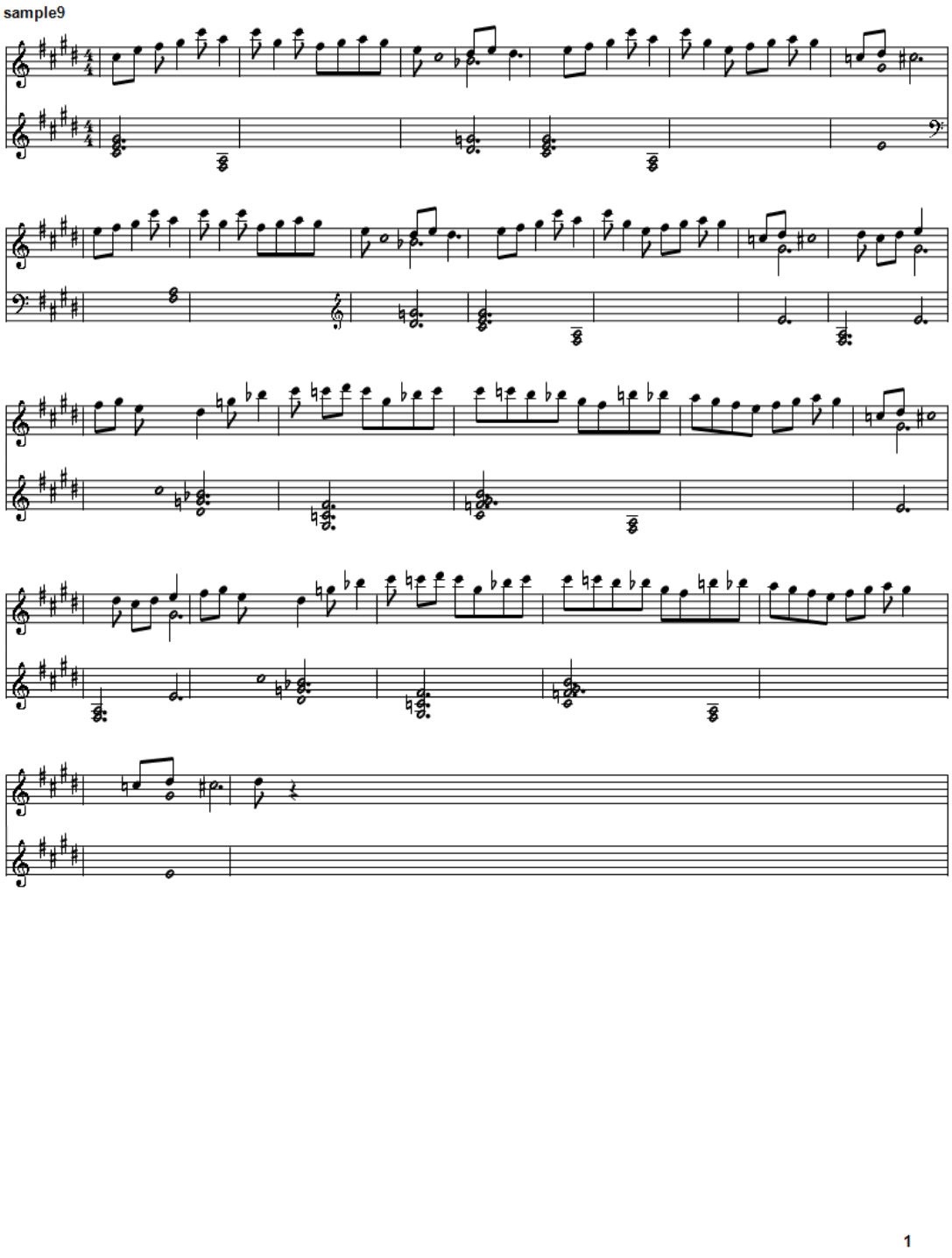

図 5.5: A music in teacher data which has the highest similarity with figure 5.4.

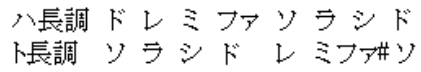

図 5.6: Comparison each note's function in C-major and G-major.

ソド (ハ長調) とレシレソ (ト長調) という二つのメロディーがあった時，相対音感者はメロディーを機能に基づいて判断するため，両者は同一視される．(図 5.6 に示される音は上下で同じ機能を持つ．)

本研究では，自動作曲システムの創作物・教師データ間の類似度を計算したが，生成された曲が音楽的に優れているかどうかについては判断しなかった．今後の課題として，生成された曲を音楽的に評価する指標が求められる．

第 6章 結論

以上本研究では，自動作曲システムの構築及びその評価法の検討を行った．自動作曲システムを LSTM を用いて作り，教師データに前処理を行うことによって次元の削減を行った結果，最高で教師データと 70% 程度一致するような生成物が得られるような，学習精度の高いニューラルネットワークが得られた．一方で創作物は少なくとも,「既存の創作物をある程度参考にしている.」「既存の創作物の完全な模倣とはなっていない」という二つの条件を満たすべきである．この二つの条件を両方完全に満たすことは出来ないため，ある指標があり，その数値が中庸であることが望ましい，というような指標があればよい．以上を踏まえて自動作曲システムによって生成された音楽が創作物として適切であるかを評価するため，類似度という指標を導入した．その妥当性をアンケートを取り，人間の感覚とどの程度一致するかを調査した．その結果，本研究によって使用した類似度 D は人間の感覚と相関係数 0.92 の強い相関があった．この類似度を用いることによって，自動作曲システムによって生成された音楽の一部は創作物として適切であると評価できた．即ち，この自動作曲システムは教師データの学習を行い，教師データの模倣能力を持つ上で作品の創作が出来ると結論付けられる．

謝辞

本研究を進めるに当たって指導教官として丁寧に様々なご指導を頂きました伊庭教授，親身に指導を頂きました長谷川先生に感謝の意を表します．また，計算機関連でご指導いただいた土肥助教授，藤田先輩，研究に当たっての助言を頂いた成瀬君を筆頭とした伊庭・長谷川研究室の皆様，学問の枠に囚われない議論を交わした高校の同期生皆様，貴重な時間を割いてアンケート調査に協力して下さった方々に感謝させて頂きます．最後に，これまで支えてくださった家族に感謝致します．ありがとうございました．

参考文献

[1]Lee, Honglak, et al. "Unsupervised feature learning for audio classification using convolutional deep belief networks." Advances in neural information processing systems. 2009.

[2]Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35.8 (2013): 1798-1828.

[3]Le, Quoc V. "Building high-level features using large scale unsupervised learning." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[4] 栗田 多喜夫, 麻生 英樹, 梅山 伸二, 赤穂 昭太郎, 細美 章隆, "多層パーセプトロンの学習における中間層に付加したノイズの影響とネットワークの可視化", 電子情報通信学会論文誌. D-II, 情報・システム, II-情報処理 J79-D-2(2), 257-266, 1996-02-25

[5]Masashi Sekino, Katsumi Nitta, "Rank Reduction by Cross-Validated Backpropagation", 電子情報通信学会技術研究報告. NC, ニューロコンピューティング 107(410), 31-36, 2007-12-15

[6]Liu, I., and Bhiksha Ramakrishnan. "Bach in 2014: Music Composition with Recurrent Neural Network." arXiv preprint arXiv:1412.3191 (2014).

[7]Eck, Douglas, and Juergen Schmidhuber. "A first look at music composition using lstm recurrent neural networks." Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale (2002).

[8]Pineda, Fernando J. "Generalization of back-propagation to recurrent neural networks." Physical review letters 59.19 (1987): 2229.

[9]Jaeger, Herbert. Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach. GMD-Forschungszentrum Informationstechnik, 2002.

[10]Gers, Felix A., Jurgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." Neural computation 12.10 (2000): 2451-2471.

[11]Ando, Daichi, et al. "Computer Aided Composition for Contemporary Classical

Music by means of Interactive GP." The Journal of the Society for Art and Science (2005).

[12]Riedmiller, Martin, and Heinrich Braun. "A direct adaptive method for faster backpropagation learning: The RPROP algorithm." Neural Networks, 1993., IEEE International Conference on. IEEE, 1993.

[13]Sonderby, Soren Kaae, et al. "Convolutional LSTM Networks for Subcellular Localization of Proteins." arXiv preprint arXiv:1503.01919 (2015).

[14]Boulanger-Lewandowski, Nicolas, Yoshua Bengio, and Pascal Vincent. "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription." arXiv preprint arXiv:1206.6392 (2012).

[15]Schulz, Hannes, and Sven Behnke. "Object-class segmentation using deep convolutional neural networks." Proceedings of the DAGM Workshop on New Challenges in Neural Computation. 2011.

[16]Fukushima, Kunihiko. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." Biological cybernetics 36.4 (1980): 193-202.

[17]LeCun, Yann, Koray Kavukcuoglu, and Clement Farabet. "Convolutional networks and applications in vision." Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. IEEE, 2010.

[18]Bay, Mert, Andreas F. Ehmann, and J. Stephen Downie. "Evaluation of Multiple-F0 Estimation and Tracking Systems." ISMIR. 2009.

[19]Eck, Douglas, and Juergen Schmidhuber. "A first look at music composition using lstm recurrent neural networks." Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale (2002).

[20] 村原京子, and 田中京子. "GF ヘンデルと JS バッハ-作品における互いの影響." 鹿児島大学教育学部研究紀要. 人文・社会科学編 48 (1997): 49-56.

[21] 二橋, and 潤一. "シューマンの和声法におけるダイナミズム-クライスレリアーナの分析を通して." (2009).

[22]Ng, Andrew. "Sparse autoencoder." CS294A Lecture notes (2011): 72.

[23]Bengio, Yoshua, et al. "Greedy layer-wise training of deep networks." Advances in neural information processing systems 19 (2007): 153.

[24]Hinton, Geoffrey, Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.

[25] 加藤充美. "MIDI 規格誕生の背景と規格の概要: 電子音楽をとりまく環境の変化 (¡ 小特集¿ MIDI 規格がもたらしたものと今後の展望)." 日本音響学会誌 64.3 (2008): 158-163.

[26] 中野倫靖, 吉井和佳, and 後藤真孝. "確率的生成モデルに基づく音楽の類似度とあり

がち度の推定に関する検討.” 情報処理学会 研究報告 音楽情報科学研究会 (2014): 1-7.

[27]Fujishima, Takuya. ”Realtime chord recognition of musical sound: A system using common lisp music.” Proc. ICMC. Vol. 1999. 1999.

[28]Newman, Mark. Networks: an introduction. Oxford University Press, 2010.

[29]Smith, Temple F., and Michael S. Waterman. ”Identification of common molecular subsequences.” Journal of molecular biology 147.1 (1981): 195-197.

[30] 深山覚, et al. ”Orpheus: 歌詞の韻律に基づいた自動作曲システム.” 情報処理学会研究報告 (2008): 179-184.

[31] 峯松信明, et al. ”音声に含まれる言語的情報を非言語的情報から音響的に分離して抽出する手法の提案—人間らしい音声情報処理の実現に向けた一検討—.” 電子情報通信学会論文誌 D 94.1 (2011): 12-26.

付録

和音及び音形の組み合わせ

和音とは 88 音の組み合わせであるから，各音が鳴っている，鳴っていないの二通りの状態を取る時 2^{88} 通りの組み合わせが存在する．また，音形は位置シフトを同一視することを利用して，全ての音が鳴っていない状態である時を除き，最も低い音が鳴っている状態と同一視できる．そのため，残りの 87 音が取りうる状態は 2^{87} 通り，これに全ての音が鳴っていない状態を加えて $2^{87}+1$ 通りとなる．