2005－MUS－59　(1)
2005／2／18

ユーザに専門知識を要求しない自動作曲システムの研究

小寺　慶生
東京工科大学大学院メディア学研究科

本研究では、ユーザに音楽の専門知識を要求せずに、ユーザが必要とする曲（フレーズ）を自動作曲できるシステムの構築を試みる。音楽の専門知識を要求しないため、システムとユーザとのインタラクションにおいて、ユーザは形容詞を用いて音楽の意味内容を指定する。ユーザはこれらの形容詞を用いて、自動作曲システムの提示する曲を評価する作業を繰り返す。この自動作曲システムは、ニューラルネットワークを用いて、曲とユーザの形容詞による評価の対を学習し、ユーザの指定した意味内容を持つ曲を自動作曲できるようになる。システムのプログラミングには Java 言語を使用した。また、ユーザに対する結果の提示には MIDI ファイルを利用する。このシステムはゲーム・プランナーやウェブページ・クリエイターなど音楽家ではないが、音楽の提供を必要とする人たちに役立つと考えられる。

A System of Algorithmic Composition for Non-Musicians

Keiki Kodera
Tokyo University of Technology, Graduate School of Media Science

The present thesis aims at building a system of algorithmic composition for non-musicians. To produce a piece of music with the system, the user is required to specify characteristics of the piece not in music-theoretical terms but in everyday language, or common adjectives such as "bright," "happy," and "depressing," and train the system until it produces a piece meanings of which the user designates by the adjectives. In other words, the system does not presuppose any categories of musical meanings. Instead, the categorization is accomplished by the neural nets implemented in the system. The system is written in Java and produces MIDI files. The system seems useful for video game planners and Web pages creators who are not musicians but need to provide music.

はじめに

本研究では、ユーザに音楽の専門知識を要求せずに、ユーザが必要とする曲を自動作曲できるシステムの構築を試みる。言い換えれば、ユーザが音楽の意味内容を指定することにより、ユーザが必要な曲を自動作曲できるシステムの構築を試みるということである。

従来、多く研究・開発されてきた自動作曲システムでは、そのシステムの開発者が音楽様式を分類し、ユーザはその分類の中から音楽様式を選んだ。しかし、そのシステムの開発者とユ

ーザの音楽様式の分類の仕方に違いが生じることは、しばしば起こることである。本研究では、システムの開発者が、あらかじめ音楽様式を分類しておくのではなく、ユーザに様式を分類する作業を委ねる。しかし、音楽を構成する要素をパラメータ化して扱った場合、分類する作業において、ユーザに音楽の専門知識の理解を要求せざるを得なくなる。

ユーザに、音楽の専門知識や専門用語の理解を要求しないために、本研究では、音楽を構成するいくつかの要素をパラメータ化して、そのパラメータをユーザに決定させるという方法を採らない。ユーザは、意図する曲の意味内容を、いくつかの形容詞で表現する。言い換えれば、音楽の意味内容を指定するということである。用いる形容詞は、音楽の専門用語である必要はない。例えば、「明るい」、「楽しげな」、「派手な」といった日常的に使われるものでよい。これらのような形容詞を用いて、ユーザは自動作曲システムの提示する曲を評価する作業を繰り返す。この自動作曲システムは、ニューラルネットワークを用いて、曲とユーザの形容詞による評価の対を学習する。つまり、ユーザの指定する意味内容によって、システムが音楽を分類の仕方を学習するのである。

本研究では、上述の自動作曲システムのフレーズ生成に焦点を絞って研究を行った。

ニューラルネットと自動作曲

ニューラルネットワークとは、人間の脳の構造を模倣して作った情報処理機構のことで、1943 年に McCulloch と Pitts らにより研究が始まり、以来、多くの分野でニューラルネットワークは利用されてきた。ニューラルネットワークを用いた自動作曲の例には、Dolson(1991)がある。Dolson は、ニューラルネットに良いリズムパターン 12 個、悪いリズムパターン 28 個を学習させて、リズムパターンの良し悪しを評価させた。また、Nishijima と Watanabe (1992)では、あらかじめフレーズ学習したシステムと人間の演奏者によるジャムセッションの研究を行った。

自動生成するフレーズの条件

Dolson(1991)では、リズムパターンに、次のような条件を定めた。

- 4/4 拍子
- 1 小節
- 八分音符以下の音価は現れない

これによりフレーズ中のタイムポイントの数は 8 個となる。Dolson はニューラルネットの入力をリズムパターンにしたため、この 8 個というのがそのまま、入力ユニット数になる。

本研究では、リズムパターンの評価ではなく、フレーズの評価を行う。このため、まず、フレーズの条件を以下のように定めた。

- 4/4 拍子
- 1 小節
- 八分音符以下の音価は現れない
- ピッチは 12 平均律

試行１：１つのニューラルネットワークによるフレーズの学習

本研究ではいくつかの試行を行ったが、本論文においては、その中から 2 つの試行を取り上げる。1 つ目の試行では、1 つのニューラルネットワークを用いて、アタックの情報とピッチの情報の両方を学習させた。

上記の条件の下で、フレーズ中に存在するタイムポイントの数は、

2 小節×8 個＝16 個

となる。この 16 個のタイムポイントがそれぞれ、アタックとピッチの情報を持つ。このピッチとアタックの情報を以下のように、0 から 31 の 32 個の符号で表した。

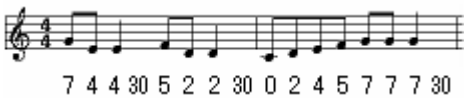
- タイムポイントに音符のアタックが存在する場合、その音符のピッチを 0 から 28 の符号で表す。
- タイムポイントが休符（無音）の場合、29 で表す。
- タイムポイントに前のタイムポイントからの音符の継続がある場合、30 で表す。
- 31 はプログラム上の例外処理に使うので、実際には意味を持たない。

32 個（5bit）の符号を扱うので、１つのタイムポイントにつき、5 個の入力ユニットが必要になる。タイムポイントは 16 個あるので、入力ユニット数は、

5 個×16 個＝80 個

となる。

以下に、具体例を示す。



出力ユニット数は、ユーザがフレーズの評価に用いた形容詞の数に依存するが、この試行１では、形容詞の数は２個と仮定した。中間層の総数は１つとして、その中間層のユニット数は８個とした。

試行１で生じた問題

このシステムを用いて、200 回以上のインタラクションを行ったが、十分と思われる結果は得られなかった。このニューラルネットワークの持つ結合重み変数の個数は、

80×8＋8×2＝656 個

である。インタラクションの回数は、ネットワークの学習データの個数に相当するが、656 個の結合重み変数を持つニューラルネットワークに対して、200 個の学習データでは十分でないことは明らかである。しかし、ユーザにこれ以上のインタラクションの回数を要求することは現実的でないと考えたため、ニューラルネットワークの規模を小さくする検討の必要性が生じた。ニューラルネットワークの規模を小さくするためには、同時に、学習データ、つまり、フレーズの表現の仕方も変える必要性がある。

試行２：複数のニューラルネットワークを用いたフレーズ学習

試行１で生じた問題を解決するため、小さなニューラルネットワークを複数用意し、それらが音楽的要素を分担して扱うように変更した。複数のニューラルネットワークにそれぞれ音楽的要素を分担させるというこの発想は、Nishijima と Watanabe（1992）を参考にした。また、扱う音楽的要素は Cope（1991）の第 4 章の音楽様式を明らかにするパラメータのリストを参考にした。

試行１と同様に、試行２で用いる形容詞の数は２個と仮定した。試行２では、４つのニューラルネットワークを用いた。この４つのネットワークは以下のとおりである。

- ネットワーク１
  - 拍点のピッチを学習する
  - 各拍点のピッチを 3 bit で表して入力として、学習は２小節単位で行う
  - 入力ユニット数：3bit×8 拍=24
  - 中間層のユニット数：4
  - 重み変数の個数：24×4＋4×2=104
- ネットワーク２
  - 拍点と裏拍の関係を学習する

- 拍点と裏拍のピッチをそれぞれ 3 bit で表し入力として、学習は 1 小節単位で行う
- 入力ユニット数：3bit×8 個=24
- 中間層のユニット数：8
- 重み変数の個数：24×8＋8×2=208

• ネットワーク 3
- ピッチとアタックの拍点での関係を学習する
- ピッチは 3 bit、アタックは 1 bit で表し、学習は 2 小節単位で行う
- 入力ユニット数:3bit×8 拍＋1bit×8 拍=32
- 中間層のユニット数： 4
- 重み変数の個数：32×4＋4×2=136

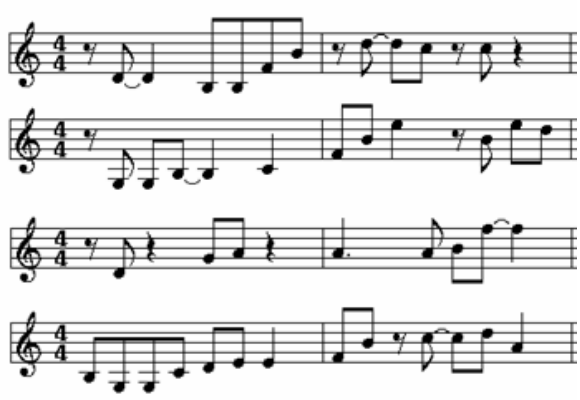
• ネットワーク 4
- 各タイムポイントのアタックを学習する
- アタックは 1 bit で表し、学習は 2 小節単位で行う
- 入力ユニット数： 1 bit×16 個=16
- 中間層のユニット数： 4
- 重み変数の個数：16×4＋4×2＝72

ピッチの表し方は、試行 1 では 5 bit で表現していたが、ネットワークを小さくするために、3 bit に変更し、扱うことのできるピッチの数は 8 個となった(プログラムの例外処理に 1 つの符号を充てるため、実質的には 7 個)。

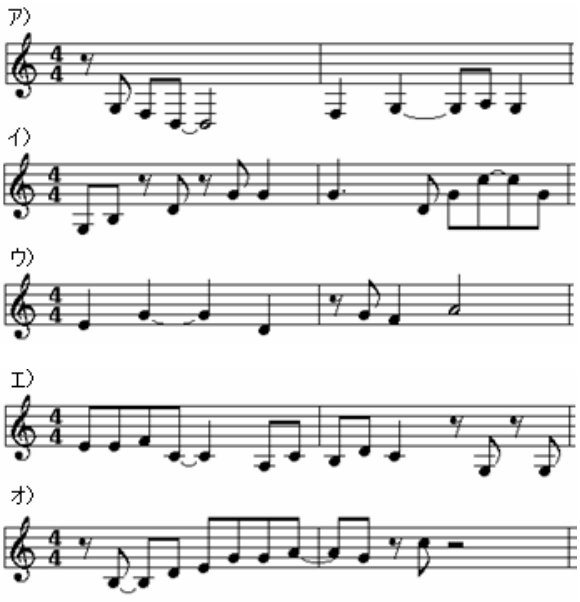
この自動作曲システムの作業の過程は、「フレーズの学習過程」と「フレーズの生成過程」がある。「フレーズの学習過程」ではシステムが生成するフレーズをユーザが形容詞をもいいて評価する。この作業を何度も程度繰り返す。「フレーズの生成過程」では、学習済みのニューラルネットワークをフィルタとして用いる。乱数でフレーズを作り、このフィルタを通過できるフレーズができるまで、その乱数によるフレーズ生成を繰り返し、フィルタを通過できたらそのフレーズをユーザに示す。

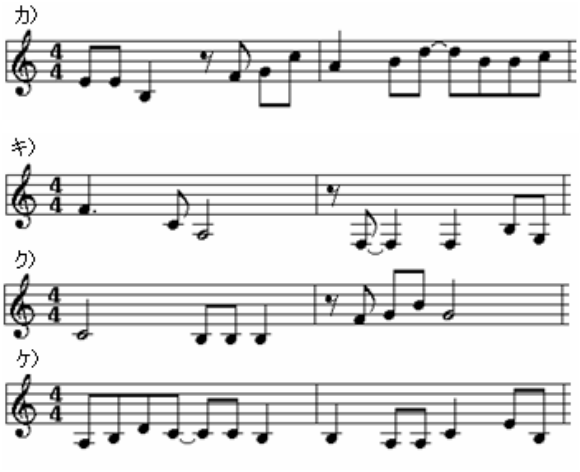
試行２の結果と考察

フレーズの学習過程において、インタラクションは 105 回行った。フレーズの生成は、本来、フレーズの学習過程が終わってから行うことを想定しているが、学習過程の経過を調べるために、学習過程の途中で、あえて、フレーズの生成を数回試みた。学習過程の途中で試みたフレーズ生成の結果を以下に示す。上から順に 45 回、65 回、75 回、83 回の学習終了時に、試みたフレーズ生成の結果である。



次に、フレーズの学習過程の終了後（105 回のインタラクションの終了後）に行ったフレーズの生成結果を示す。





試行２では、１０５回のインタラクションを行った。フレーズの評価に用いた形容詞は、「明るい」と「楽しい」の２つである。また、初期学習データとして、インタラクションの開始前に３１個のフレーズを与えた。したがって、学習データの数は 105＋31＝136 個である。

４つのニューラルネットワークのうち、最も多くの重み変数を持つのは、ネットワーク２であるが、このネットワークは学習の単位を１小節としたため、１つのフレーズから２つの学習データを得ることができる。この点を考慮したとき、最も多くの学習用フレーズを必要とするのは、２番目に多くの重み変数を持つ、ピッチとアタックの拍点での関係を学習するネットワークになる。このネットワークの重み変数の数は 136 個である。結合重み変数が 136 個のニューラルネットワークに対して 136 個の学習データというのは、十分な量とは言いがたいが、試行１と比較すれば、大幅な改善といえる。

しかし、105 回のインタラクションには、ニューラルネットの学習に要した時間を含めて約 11 時間を要した（動作環境は PentiumIII 1GHz、メモリ 256MB、WindowsXP）。試行１から試行２への変更の目的はユーザの負担を軽減することであり、そのために必要なインタラクションの回数を減らすための変更を行った。試行１から試行 2 への変更を簡単に言えば、すべての要素を取り扱う１つの大きなニューラルネットワークを、限定された要素を取り扱う４つ小さなニューラルネットワークに分割した。その結果、確かにインタラクションの回数は減ったが、1 回のインタラクションに要する時間が増大した。

次に、生成されたフレーズについて考察を行う。フレーズの生成過程の途中で生成したフレーズと、学習過程が終了した後で生成したフレーズを比較すると、生成過程の比較的初期の段階において生成したフレーズに、不自然な跳躍（45 回目終了時のフレーズの第 1 小節後半や 65 回目終了時のフレーズの第 2 小節の前半など）が見られるが、学習過程終了後の生成結果では、不自然な跳躍が減っていることが分かる。

フレーズの生成過程の途中段階と学習終了後のフレーズ生成の比較から、フレーズの含む要素すべてではないにしろ、フレーズを構成する一部の要素と形容詞の関係は学習され、生成結果に反映されたと考えられる。学習終了後に生成したフレーズは、単に不自然な跳躍が減っただけでなく、学習初期段階に生成したフレーズと比較してみると、より自然なフレーズに聴こえる。

しかし、終了後に生成したフレーズすべてが、形容詞の評価を満たしているとは言い切れない。例えば、（聴こえ方に個人差があるため断定できないが）終了後の結果（ケ）は、明るくないと感じる人も多いかもしれない。

この原因の１つとして、フレーズの生成過程におけるインタラクションの回数の不足が考えられる。インタラクションの回数を増やせば、より良い結果が期待できる。しかし、全 105 回のインタラクションに要した時間を考えると、回数を増やすことは難しい。

おわりに

従来、多く行われてきた音楽様式を指定する自動作曲とは違い、フレーズの意味内容を指定することにより、ユーザの必要なフレーズを自動作曲するシステムを作成した。評価に用いる形容詞を変えたらどうなるのかを検証していないが、生成結果からある程度、意味内容が反映されることが確認できた。

また、問題点も明らかになった。インタラクションにかかる時間が大きいことである。本研究で用いた方法では、ニューラルネットワークの学習データを多く用意するために、インタラクションの回数は多いほうが望ましいが、インタラクションに要する時間が長くなってしまい、ユーザの負担が大きくなってしまうことが分かった。

本研究ではフレーズ部分に着目したため、曲の構成やハーモニーなどを扱っていないが、これらを含めた、曲を自動生成するシステムを構築する際に、本研究で試みが応用可能だと考えられる。また、別の応用方法として、本システム自体を、例えば、ゲームコンテンツに埋め込んで、ゲーム中のインタラクションと関連付けて利用すれば、そのゲームの音楽において面白い効果が期待できる。

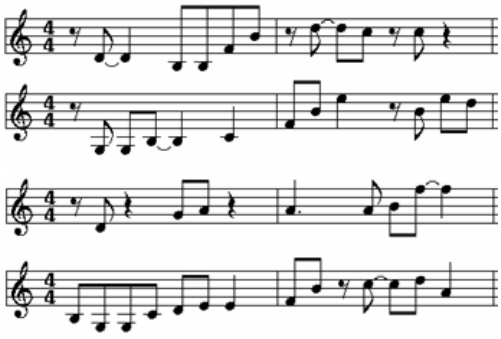
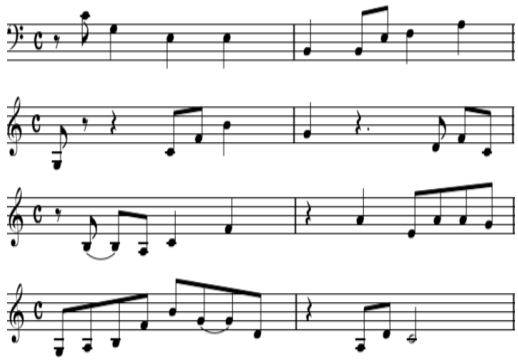
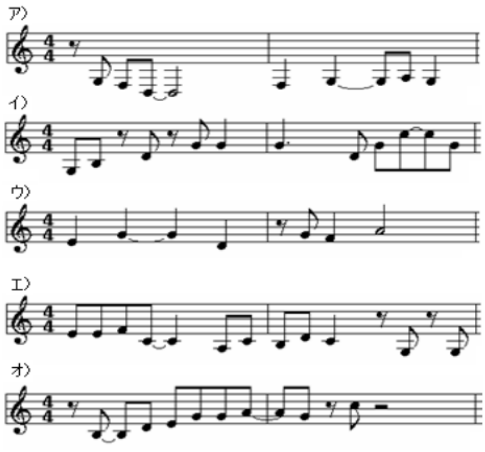
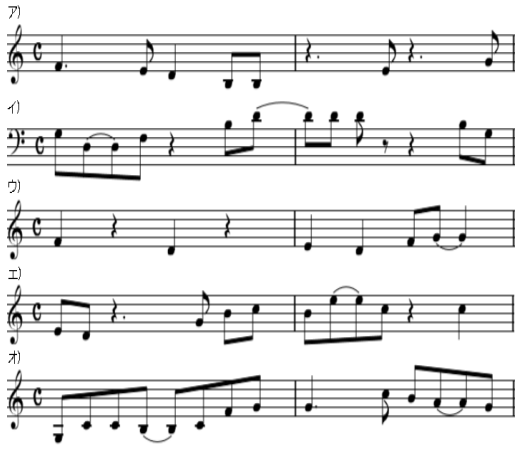
今後解決すべき問題として、１回のインタラクションにかかる時間の短縮が必要である。１回のインタラクションの時間が短縮されれば、インタラクションの回数を増やすことも可能になり、出来上がるフレーズの質がより良くなることが期待できる。４つのニューラルネットワークのうち、学習に最も時間がかかるのは、拍点と裏拍の関係を学習するニューラルネットワークである。この学習を効率的にするための１つの方法として、和声的な構造を扱う要素に加えることが考えられる。拍点と裏拍のピッチの関係は、和声音・非和声音と関係するため、和声構造を扱うニューラルネットワークを導入することで、問題が改善できる。本研究では多層パーセプトロンのみを利用したが、他のニューラルネットワークを試すことで改善できる見込みもある。

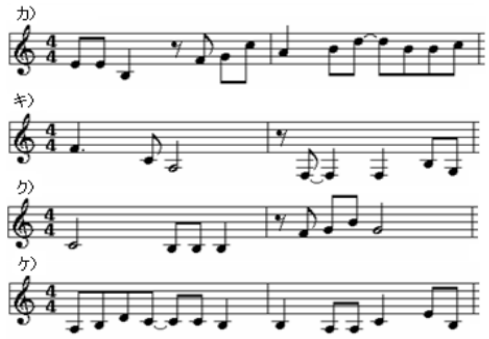
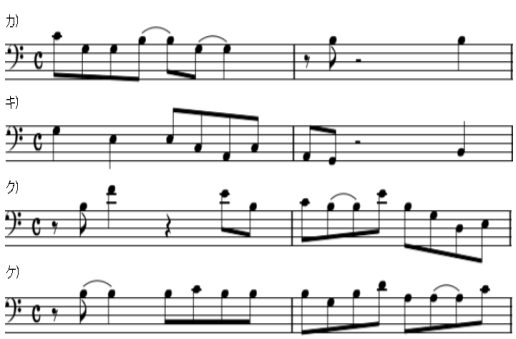
参考文献

1) Dolson, Mark. “Machine Tongues XII: Neural Networks.” *Music and Connectionism*, Ed. by Peter M. Todd and D. Gareth Loy. Cambridge, Massachusetts: The MIT Press, 1991: 3-19.
2) Nishijima, Masako and Kazuyuki Watanabe. “Interactive music composer based on neural networks.” *International Computer Music Conference Proceedings*, San Jose, California: The International Computer Music Association, 1992: 53-56.
3) Cope, David. *Computers and Musical Style.* Madison, Wisconsin: A-R Editions, Inc, 1991.
4) Curtis Roads（青柳龍也（他）訳）『コンピュータ音楽』（歴史・テクノロジー・アート）東京電機大学出版局、2001。
5) 甘利俊一、酒田英夫 編『脳とニューラルネット』朝倉書店 1994。
6) Russell Beals and Tom Jackson（八名和夫監訳）『ニューラルコンピューティング入門』海文堂 1993。
7) jMusic. http://jmusic.ci.qut.edu.au/, 2004.7.12 取得。
8) 静岡理工科大学情報システム学科菅沼研究室 http://www.sist.jp/~suganuma/index.html、2002.12.13 取得。
9) Dodge, Charles and Thomas A. Jerse, *Computer Music*. New York: Schirmer Books, 1997.
10) Quine, W.V.O. “Natural Kinds.” *Ontological Relativity and Other Essays.* New York: University Press, 1969: 114-138.

正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

訂正箇所	誤	正
2 ページ 左側 27 行目の後に追記	本研究では、上述の自動作曲システ ムのフレーズ生成に焦点を絞って研究 を行った。	本研究では、上述の自動作曲システ ムのフレーズ生成に焦点を絞って研究 を行った。自動作曲システムのニュー ラルネットワーク部分は 6)および 8) を参考にコンピュータプログラムを作成した。
2 ページ 左側 29 行目	ニューラルネットワークとは、人間の脳の構造を模倣して作った情報処理機構のことで、	ニューラルネットワークとは、脳の作りの一部をコンピュータにおける情報処理に組み込んだもののことで、
4 ページ 右側 9 行目 楽譜の説明と楽譜	上から順に 45 回、65 回、75 回、83 回の学習終了時に、試みたフレーズ生成の結果である。 	上から順に 58 回、76 回、79 回、95 回の学習終了時に、試みたフレーズ生成の結果である。 
4 ページ右下 楽譜 ア）～オ）		

5ページ左上 楽譜 カ）～ケ）		
5ページ左側 4行目	初期学習データとして、インタラクションの開始前に３１個のフレーズを与えた。したがって、学習データの数は105＋31＝136個である。	初期学習データとして、インタラクションの開始前に6個のフレーズを与えた。したがって、学習データの数は105＋6＝111個である。
5ページ左側 16行目	結合重み変数が136個のニューラルネットワークに対して136個の学習データというのは、十分な量とは言いがたいが	結合重み変数が136個のニューラルネットワークに対して111個の学習データというのは、十分な量とは言いがたいが
5ページ右側 13行目	45回目終了時のフレーズの第1小節後半や65回目終了時のフレーズの第2小節の前半など	76回目終了時のフレーズの第1小節など
6ページ左側 7行目	生成結果からある程度、意味内容が反映されることが確認できた。	生成結果から、意味内容が十分に反映されたとはいえない。
6ページ右側 参考文献	1) Dolson, Mark. “Machine Tongues XII: Neural Networks.” *Music and Connectionism*, Ed. by Peter M. Todd and D. Gareth Loy. Cambridge, Massachusetts: The MIT Press, 1991: 3-19. 2) Nishijima, Masako and Kazuyuki Watanabe. “Interactive music composer based on neural networks.” *International Computer Music Conference Proceedings*, San Jose, California: The International Computer Music Association, 1992: 53-56. 3) Cope, David. *Computers and Musical Style.* Madison, Wisconsin: A-R Editions, Inc, 1991. 4) Curtis Roads（青柳龍也（他）訳）『コンピュータ音楽』（歴史・テクノロジー・アート）東京電機大学出版局、2001。 5) 甘利俊一、酒田英夫 編『脳とニューラルネット』朝倉書店 1994。 6) Russell Beals and Tom Jackson（八名和夫監訳）『ニューラルコンピューティング入門』海文堂 1993。 7) jMusic. http://jmusic.ci.qut.edu.au/, 2004.7.12取得。 8) 静岡理工科大学情報システム学科菅沼研究室 http://www.sist.jp/~suganuma/index.html 、2002.12.13取得。 9) Dodge, Charles and Thomas A. Jerse, *Computer Music*. New York: Schirmer Books, 1997. 10) Quine, W.V.O. “Natural Kinds.” *Ontological Relativity and Other Essays.* New York: University Press, 1969: 114-138.	1) Dolson, Mark. "Machine Tongues XII: Neural Networks". Music and Connectionism. Todd, Peter M.; Gareth Loy, D., eds. Cambridge, Massachusetts, The MIT Press, 1991, p.3-19. 2) Nishijima, Masako.; Kazuyuki, Watanabe. "Interactive music composer based on neural networks". International Computer Music Conference Proceedings, San Jose, California, The International Computer Music Association, 1992, p.53-56. 3) Cope, David. Computers and Musical Style. Madison, Wisconsin, A-R Editions, Inc, 1991. 4) Roads, Curtis. コンピュータ音楽:歴史・テクノロジー・アート. 青柳龍也ほか訳, 東京電機大学出版局, 2001. 5) 甘利俊一, 酒田英夫編. 脳とニューラルネット. 朝倉書店. 1994, p.1-14. 6) Beale, Russell.; Jackson, Tom. ニューラルコンピューティング入門, 八名和夫監訳, 海文堂, 1993. p.1-89. 7) Sorensen, Andrew.; Brown, Andrew. jMusic. http://jmusic.ci.qut.edu.au/, 2004.7.12取得. 8) 静岡理工科大学情報システム学科菅沼研究室, http://www.sist.jp/~suganuma/index.html, 2002.12.13取得. 9) Dodge, Charles.; Jerse, Thomas A., Computer Music. New York, Schirmer Books, 1997. 10) Quine, W.V.O. "Natural Kinds". Ontological Relativity and Other Essays. New York, University Press, 1969, p.114-138. 11) Iwata, Akira.; Matubara, Toshiyuki. ニューラルネットワーク入門. 1996. http://mars.elcom.nitech.ac.jp/java-cai/neuro/menu.html, 2002.6.10取得. 12) SoftComputing lab. ニューラルネットワーク用語集.

http://kyu.pobox.ne.jp/softcomputing/neuro/words.html, 2002.12.13 取得.
13) Udemy. "ニューラルネットワークとは？人工知能の基本を初心者向けに解説！". Udemy メディア. https://udemy.benesse.co.jp/ai/neural-network.html, 2018.4.28 取得.