平成 26 年度 修士論文

MIDI ギターと NMF を用いた音響信号処理の統合によるギター演奏の自動採譜

指導教員 北原鉄朗准教授

日本大学大学院総合基礎科学研究科

6113M03 大塚匡紀

2015 年 2 月 提出

概　　要

自動採譜とは、コンピュータによって楽器演奏を楽譜やMIDIデータなどと言ったフォーマットに変換することを指す。自動採譜は、音楽学習や演奏支援などの分野において需要があり、音楽情報処理において重要な課題の一つである。数ある楽器のなかでもギターは、ポピュラー音楽で頻繁に使用され、また演奏人口も多い。そのため、ギター演奏を対象とした自動採譜は、大きな需要があり、重要な課題の1つである。

ギター演奏をリアルタイムで自動採譜する機材に、MIDIギターが存在する。MIDIギターは、コンピュータを用いて音楽制作を行う際に、ギタリストが自分の慣れた楽器を使ってデータ入力ができるといった点において有用であるが、弦の振動をピックアップで取得する構造のため、ピッキングの取りこぼしなどが発生し、MIDIキーボードに比べると入力される演奏情報の正確性に難がある。特に、ファンクなどのジャンルのコードカッティングの演奏において、3つの問題が頻繁に発生する。それらの問題は、存在する音の欠落、存在しない音の採譜、連続する短い音の採譜である。

そこで本研究では、MIDIギターピックアップによる処理と非負値行列因子分解(Non-Negative Matrix Farcrization)による音響信号処理を統合することによって、3つの問題点を解決し、採譜の高精度化を目指す。まず、第3章では、音響信号処理のみによる自動採譜手法について検討する。MIDIギターと同様の用途で使用可能なようにリアルタイムでの採譜を視野に入れ、音響信号処理は随時的な処理(オンライン・アルゴリズム)にて採譜する。しかし、本来のNMFはオンラインでは

採譜することができない。そこで、予備演奏を用いることでオンラインにに採譜できるように改良する。その手法では、単純な閾値処理により発音を検出するが、演奏者のくせや倍音などの要因により閾値決定が難しい問題がある。そこで、新たに予備演奏を加えニューラルネットワークを用いて弦ごとに閾値調整を行い、発音検出を行えるように改良する。単純な閾値処理による採譜手法とニューラルネットワークを用いた発音検出手法で各々の手法で評価実験を行ったところ、単純な閾値処理において様々な閾値を試して最も良かった結果とニューラルネットワークを用いた発音検出手法の結果が同等のものとなった。次に、第 4 章では MIDI ギターと 3 章で提案した手法とをオンラインに統合する手法を述べる。MIDI ギターと NMF による音響信号処理手法で採譜の傾向に違いが見られたため、傾向を考慮した統合により MIDI ギターの問題点である存在する音の欠落、存在しない音の採譜、連続する短い音の採譜を解決する。統合にあたり、積による統合と和による統合、ニューラルネットワークによる統合の 3 種類を提案する。それらの手法について評価実験したところ、提案手法が MIDI ギターよりも平均の再現率が 0.067、適合率が 0.402、F 値が 0.279 向上した。第 5 章にて、今後の課題や展望について述べる。そして、第 6 章にて結論を述べる。

目　　次

図 目 次

表 目 次

第1章　序　　論

1.1　背景

自動採譜とは、コンピュータによって楽器演奏の音響信号を楽譜、タブ譜、MIDI データといった演奏内容を記述したフォーマットに変換することである。本研究では、その中でも MIDI データ [1] のフォーマットを採用する。MIDI は、Musical Instrument Digital Interface のことであり、コンピュータが解読可能な演奏記述のフォーマットである。MIDI は、シンセサイザーやシーケンサといった電子楽器の演奏データを機器間で通信するための共通規格として制定された。MIDI データには、楽器演奏を表現するための三要素である音高、発音時刻、消音時刻の要素が含まれており、細かく演奏の要素を指定し、表現することができる。要するに、本研究における自動採譜とは楽器演奏から音高、発音時刻、消音時刻を抽出し、演奏内容を記述したフォーマットである MIDI データに変換するプロセスのことであり、主に楽器演奏の学習や支援などの分野において需要がある。

特にギターは、ポピュラー音楽で頻繁に使用され、演奏人口も多いため、ギター演奏を対象とした自動採譜は需要が大きい。そのため、ギター演奏を対象とした自動採譜の研究は数多く行われてきた [3]-[7]。更に、近年はコンピュータによって音楽を製作する DTM が一般に普及したことにより、ギター演奏をリアルタイムで自動変換し、MIDI 入力を行うことへの需要は高まっている。

ギター演奏をリアルタイムに自動採譜する際によく用いられる機材に MIDI ギター（図 1.2、図 1.3) がある。MIDI ギターとは、通常のギターにディバイデッドピックアップを取り付けたものであり、ギターを他の楽器の音色で演奏するシンセ

サイザーのような用途において重宝される。MIDI ギターは、ギター演奏をリアルタイムに MIDI 形式に変換して出力することができ、使い慣れた楽器で演奏情報を入力できる点において、DTM などを使って作曲するギタリストには有用である。

しかし、弦の振動をピックアップで取得するため、演奏の解析の際に、ピッキングの取りこぼしなどが発生してしまう。そのため、スイッチの構造をしている MIDI キーボードに比べると入力される演奏情報の正確さに問題がある。特に、ファンクなどで多用されるゴーストノートを含む 16 ビートのコード・カッティングにおいて頻繁に 3 つの問題が発生する。3 つの問題とは、存在する音の欠落、存在しない音の採譜、連続する短い音の融合である (図 1.1)。

図 1.1: MIDI ギターの問題の一例

1.2 目的

本研究では、非負値行列因子分解 (NMF)[12] による音響信号処理と MIDI ギターピックアップによる処理とをオンラインで統合することで、自動採譜の高精度化を目指す。統合にあたり、NMF をオンライン・アルゴリズムにて採譜できるよう

に拡張する。その拡張したNMFによる音響信号処理手法とMIDIギターとの2つの採譜処理を併用することにより採譜精度の高精度化を図る。

1.3 論文の構成

本論文の構成として、第2章では自動採譜の研究及びギター演奏を対象にした自動採譜の研究やギター演奏の商品を概観する。第3章では、NMFによるオンライン・アルゴリズムを用いた音響信号処理を述べる。まず、NMFをオンラインにて採譜できるようにする手法を提案し、更にニューラルネットワークによって採譜精度を向上させる方法を提案する。第4章では、第3章で述べた手法とMIDIギターとをオンラインにて統合する手法を述べる。第5章にて、第4章までで述べたものをまとめ、今後の課題について議論をする。第6章で結論を述べる。

図 1.2: MIDI ギター


Roland GR-55 GUITAR SYNTHESIZER
OUTPUT LEVEL
EXP PEDAL SW ON/OFF
PAGE
EDIT
SOUND STYLE
V-LINK
LEAD
RHYTHM
OTHER
USER
EZ EDIT
EXIT
ENTER
WRITE
AUDIO PLAYER
USB MEMORY
BANK
1
2
3
CTL
BANK SELECT
TUNER
REC/PLAY/DUB
PHRASE LOOP
GR
COSM

図 1.3: GR-55

第2章　ギター演奏に関する自動採譜の現状と関連する商品

本章では、まずギター演奏を対象とした研究をアプローチごとに述べる。更に、ギター演奏を対象とした商品について述べる。そして、本研究とそれらを比較し、本研究の位置づけを明らかにする。

2.1　ギター演奏を対象とした自動採譜

本節ではギターの自動採譜に特化した研究をアプローチごとに紹介する。

2.1.1　ギターの音色や身体的制約を用いたアプローチ

元々特定の楽器に限定しなかったり、ピアノを対象としていた研究をギターに特化させた研究は幾つかある。ギターという楽器の特性上、身体的な制約が顕著に表れやすいため、身体的制約を加えた様々な研究がなされている [2]-[8]。多くの研究がなされていることから現在の主流のアプローチといえる。

有元らは単音のメロディやベースの基本周波数推定手法の PreFEst[3] を改良し、ギターフォームの制約を加えるなどしギター独奏に特化させ従来の PreFEst よりも採譜精度を向上させた [2]。また、矢澤らは潜在的調波配分法 (Latent Harmonic Allocation:LHA) に押弦制約 (発音可能音域や同時発音可能な押さえ方) を加えるなどしギターに特化させ、採譜精度を向上させた [4]。これらの研究からギター

固有の押さえ方を考慮することによって採譜精度が向上することが確認できる。Barbancho らも同じように押弦制約に着目した研究を行っている [5]。

2.1.2 画像処理を用いたアプローチ

音響信号処理と画像処理を統合し、採譜精度の向上を試みた研究が幾つかある [7][6]。山上らは、演奏動画を解析した情報と MIDI ギターの情報を統合することによって高精度のギター用の楽譜を出力するシステムを開発した [7]。Paleari らは、Web カメラを用いて演奏の際の右手をリアルタイムで解析し、運指の情報と音響信号処理を統合するシステムを開発した [6]。

2.1.3 その他のアプローチ

Fiss らはギター演奏をタブ譜として採譜するシステムを構築した [8]。Fiss らのシステムは、音響信号から演奏されている音高・リズムだけでなく、弦とフレットを特定し、どのように演奏されているかも推定できる。O'Grady らは、ギター演奏の採譜精度の改良を目指し、NMF と 1 弦ごとに音響信号が取り出せるように改造した MIDI ギターを用いた研究を行っている [9]。Harquist らも、ギター演奏を対象とした NMF を用いたリアルタイムでの採譜手法を提案している [10]。また、調波音と打楽器音分離手法を用いたギター演奏の採譜の研究もある [11]。

2.2 関連商品

ハードウェア関連のものとしては、ギターの運指動作のモーションキャプチャ装置の開発 [13] やギター型の MIDI コントローラーの You Rock Guitar[14](図 2.1) やヤマハ EZ-EG[15](図 2.2) がある。これらの機器は、MIDI 変換のために設計されているため、確実に正確な MIDI 入力が可能である。しかし、これらのギター型

の MIDI コントローラーは、本来のギターの構造とはかけ離れており、ギターの演奏感覚を損なってしまうといった問題がある。

図 2.1: You Rock Guitar

図 2.2: EZ-EG

2.3 本研究の位置づけ

本研究では、演奏内容をファンクで多用されるカッティングフレーズに的を絞り、MIDI ギターと音響信号処理を統合することでギター演奏の高精度な自動採譜

の実現を目指す。音響信号処理では、アンプやエフェクタに接続するためのオーディオ出力端子から分岐させた音響信号を入力とし、エフェクタなどの影響はないものとする。MIDI ギターと音響信号処理はともにピックアップによって観測した弦の振動をもとにしているが、内部処理が異なるため誤りの傾向にも差が出てくると予想される。そこで、両者の処理結果を統合することで誤りの少ない出力が得られると期待される。上述の通り、手の大きさなどの身体的な制約を導入したり、画像などの他モダリティを併用することで精度をさらに上げることも考えられ、実際、既存研究においてそのような報告もなされているが、本研究では、このようなトップダウンな情報や他モダリティを用いない範囲での高精度化を検討する。また、MIDI ギターと同様の用途に使えるようリアルタイムでの採譜の実装を視野に入れ、逐次的な処理（オンラインアルゴリズム）のみで採譜を行う。以下、3 章では音響信号処理単独による採譜処理について検討する。その後、4 章で MIDI ギターとの統合について検討する。

第3章　NMFによる音響信号処理を用いた採譜

本章では、NMF による音響信号処理を用いた採譜手法について述べる。まず、NMF をオンラインにて採譜できるようにする手法を述べる。更に、発音検出を改善したニューラルネットワークを用いた発音検出を加えた手法を述べる。そして、各々の手法を比較した評価実験について報告する。

3.1　はじめに

本研究では、音響信号処理の手法に非負値行列因子分解:NMF(Non-Negative Matrix Factorization)[12] を用いる。NMF とは、非負値行列 V を基底行列 W と時系列の重み行列 H の積として、以下の形に近似するアルゴリズムである。

$$V \cong WH \tag{3.1}$$

NMF は音響信号処理だけにとどまらず、画像処理や自然言語処理などの分野にも応用されているアルゴリズムである。NMF を音響信号に対して適用すると多重音からの周波数推定が可能になる。実際の楽音は、個々の音の周波数成分が複雑に重なりあった多重音であり、各周波数成分がもともとどの音に由来しているのか知ることは難しい。つまり、一度足しあわされてしまった多重音を足しあわされる前の形に分解することは不良設定問題である。そこで、NMF は非負値の制約が作用し、音響信号を音高、音色、リズムといった特徴に近似的に分解することが

できる。要するに、楽器演奏をある一定の間隔でフーリエ変換した結果であるスペクトログラム V に対して適用すると音高と音色のスペクトルに相当する基底ベクトルで構成される基底行列 W と それらのスペクトルの時系列の重みである H に分解することができる。例えば、ドレミファソラシドと演奏したスペクトログラムに対して基底数8としてNMFを適用すると、ドの音のスペクトル、レの音のスペクトルのように計8音の基底ベクトルからなる基底行列 W とそれら各々ベクトルの時系列ごとの重みに分解することができる。

3.2 NMFを用いたオンラインアルゴリズムによる採譜手法

本来のNMFは、演奏が終了した信号のスペクトログラムに対して適用するものである。つまり、演奏が完了したものにのみ適用可能であり、リアルタイムの演奏を対象に採譜することはできない。本研究では、リアルタイムでの採譜を目指しているので、オンラインアルゴリズムとしてNMFを適用するために、2つのステップにより採譜する。

まず第1ステップにて、採譜したい演奏に先立ち、本演奏に用いるのと全く同じギターを用いて、各弦の各フレット (計138音) を一音ずつ順番に正確に演奏する (これを第1予備演奏とする)。この第1予備演奏に対してNMFを適用し、各弦の各フレットごとのスペクトルを表す基底ベクトルから成る基底行列 W を得る。次に第2ステップにて、採譜したい演奏 (これを本演奏とする) を演奏する。そして、第1ステップにて求めた基底行列を用いて重み行列 H を得る。重み行列から音高、発音時刻、消音時刻を求め採譜をする。

3.2.1 第 1 ステップ: 予備演奏からの基底行列の推定

はじめに、各弦の各フレットを正確に演奏し、弦 k の演奏のスペクトログラム V_k を求める (サンプリング周波数：44100Hz、窓幅：4096 点、シフト幅：10ms)。次に NMF を用いて V_k を 2 つの非負値行列 W_k 、 H_k の積 $V_k \cong W_k H_k$ に分解する。ここで W_k は弦 k における基底行列、 H_k は弦 k における重み行列を表す。基底行列 W_k は、弦 k の各フレットごとのスペクトルに相当する基底ベクトルで構成される。重み行列 H_k は、各基底ベクトルに対応するある時刻 t の重み $\boldsymbol{h}_{k,t}$ を表している。

NMF での分解の際の基底数は、各弦のフレット数すなわち実音数の 23 に対し基底数は 35 と設定した。これはフレット移動の際などに発生するノイズを考慮してのことである。ノイズを削除し、自動的に正確な基底ベクトルを推定するために、基底ベクトル $\boldsymbol{w}_k$ に対して基本周波数推定を行う。基本周波数推定は、基底ベクトル $\boldsymbol{w}_k$ の各々の値である $w_{k,n}$ がピークの時、ピーク間の周波数差を求めることによって推定する。そして、倍音の数が多いなどの特徴を持つものはノイズとみなして削除する。更に、重みベクトル $\boldsymbol{h}_k$ のうちコサイン類似度がある閾値より高い組があるとき、それらは同じ音を表していると判断し、対応する基底ベクトルを統合する。その後、分解された段階ではフレットの順に並んでいないため、各弦の各フレットの音を順番に演奏しているとの仮定の下、重みベクトルが最大になる時刻が早い順に並び替え、その順にフレットを割り当てる。

3.2.2 第 2 ステップ: 本演奏の採譜

重みベクトルの推定

本演奏に対して、フレーム (10ms) ごとに重みベクトルを求める。10ms 間隔で短時間フーリエ変換 (サンプリング周波数、窓幅は 3.2.1 節と同様) を行う。時刻

t におけるパワースペクトルを $\boldsymbol{v}_t$ とし、弦 k における基底行列 W_k の擬似逆行列 [16] を W_k^{-1} とすると、重みベクトルは以下の式で表すことができる。

$$\boldsymbol{h}_{k,t} = W_k^{-1}\boldsymbol{v}_t \tag{3.2}$$

以降、上記の式を用いて各演奏に対してフレームごとの重みベクトル $\boldsymbol{h}_t$ を求める。

MIDI形式への変換

発音・消音検出の具体的な例を図3.1に示す。弦 k の重みベクトルである $\boldsymbol{h}_{k,t}$ の各要素である重み値 $h_{k,t,n}$ が閾値 h_0 を上回ったとき、すなわち $h_{k,t-1,n} \leq h_0$ かつ $h_{k,t,n} > h_0$ を満たす時、フレット n が発音されたとみなして、フレット n に対応するノートナンバーのノートオンメッセージを出力する。同様に $h_{k,t-1,n} \geq h_0$ かつ $h_{k,t,n} < h_0$ を満たす時、フレット n に対応するノートナンバーのノートオフメッセージを出力する。なお、MIDIにおいて、ノートナンバーとは音高を表すものである。

また、同じ音高の音 n を連続してピッキングした時の発音時刻を検出するために、重みベクトル $\boldsymbol{h}_{k,t}$ の谷検出を行う。$h_{k,t-2,n} > h_{k,t-1,n} \geq h_0$ かつ $h_{k,t-1,n} < h_{k,t,n}$ を満たす時、フレット n が発音されたとみなして、フレット n を連続してピッキングしているとみなし、フレット n に対応するノートナンバーのノートオフメッセージを出力し、その直後にノートオンメッセージを出力する。

3.3 ニューラルネットワークを用いた発音検出手法

3.2節で述べた採譜手法は単純な閾値によって発音検出しているが、適切な閾値は様々な演奏の要因によって定められる。それらの要因とは、ピッキングの強さや違う弦で同じ音を表す異弦同音や倍音などである (図3.2)。更に、演奏者のくせ

図 3.1: 単純な閾値処理における発音・消音検出の例

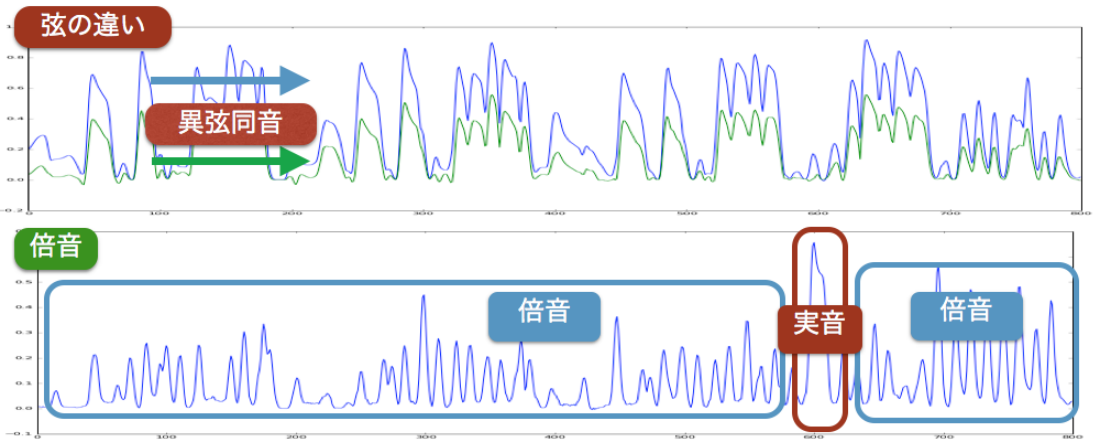

図 3.2: 閾値調整に関わる様々な要因の例

なども関わってくる。わずかな閾値の差で採譜結果には大きな差が生まれるため、それらの要因に合わせた動的な閾値調整は容易ではない。

そこで、新たに予備演奏 (これを第 2 予備演奏とする) を追加することで閾値の調整を行う。予め予備演奏の正解を与え、ユーザーの演奏と比較することで閾値の調整を行う。この閾値調整では、3.2 節の採譜手法では階段関数での発音検出だったものをシグモイド関数に置き換える。つまり、階段関数では発音か否かの 2 種類のみの判定だったが、シグモイド関数に置き換えることによって 0.0 から 1.0 の範囲の実数値から発音を判定する (図 3.4)。このケースでは、シグモイド関数に置き換えるということはニューラルネットワークで学習するのと同価であるため、図 3.3 で示したニューラルネットワークを弦ごとに用意し、それらを用いて学習をし、そのニューラルネットワークを用いて本演奏の採譜を行う。

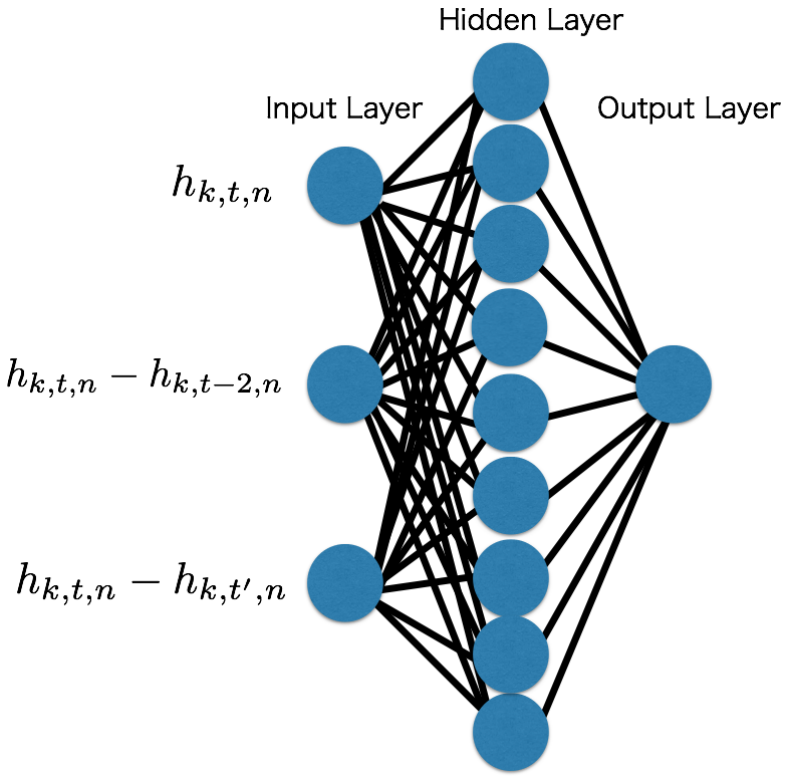

図 3.3: 学習に用いるニューラルネットワーク

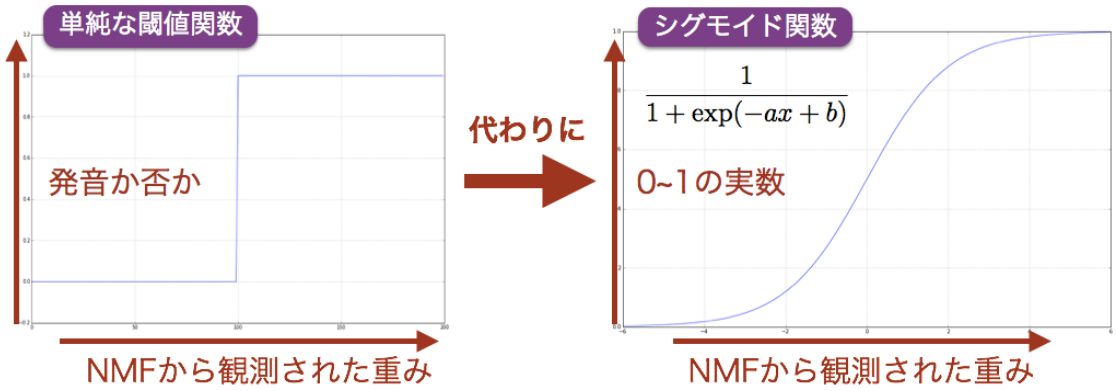

図 3.4: 階段関数からシグモイド関数への置き換え

3.3.1 第 2 予備演奏を用いた学習

3.2.2 節と同様にして、第 2 予備演奏に対して弦ごとの重みベクトル $\boldsymbol{h}_{k,t}$ を求め、$\boldsymbol{h}_{k,t}$ がピークの時、下記の処理でニューラルネットワークの学習を行う:

1. $h_{k,t,n}$ から以下の特徴ベクトル $\boldsymbol{x}_{k,t,n}$ を抽出する (図 3.5):

$$\boldsymbol{x}_{k,t,n} = (h_{k,t,n},\ \ h_{k,t,n} - h_{k,t-2,n},\ \ h_{k,t,n} - h_{k,t',n}).$$

   ここで、t' は時系列 $\{h_{k,\tau,n}; \tau = 0, \cdots, t, \cdots\}$ において時刻 t の直前の谷の時刻、すなわち、$h_{k,\tau-1,n} > h_{k,\tau,n}$ かつ $h_{k,\tau+1,n} > h_{k,\tau,n}$ を満たす τ のうち t より小さい最大値である。

2. 入力ベクトル $\boldsymbol{x}_{k,t,n}$ の対となる出力値 $s_{k,t,n}$ は以下の様に定義される:

$$s_{k,t,n} = \begin{cases} 1\ (t \text{ がフレット } n \text{ に対して発音時刻の時}) \\ 0\ (t \text{ がフレット } n \text{ に対して発音時刻でない時}) \end{cases}$$

3. $\boldsymbol{x}_{k,t,n}$ と $s_{k,t,n}$ をデータ・セットにしてニューラルネットワークの学習を行う。

学習の収束を考え、データ・セットにおける $s_{k,t,n}$ の 1 と 0 の数の比率を同じにする。なお、演奏時の僅かなずれを考慮し、時刻 t における $\boldsymbol{x}_{k,t,n}$ は、予め設定した正解の発音時刻の前後 6 フレーム以内であれば時刻 t を発音時刻とみなす。

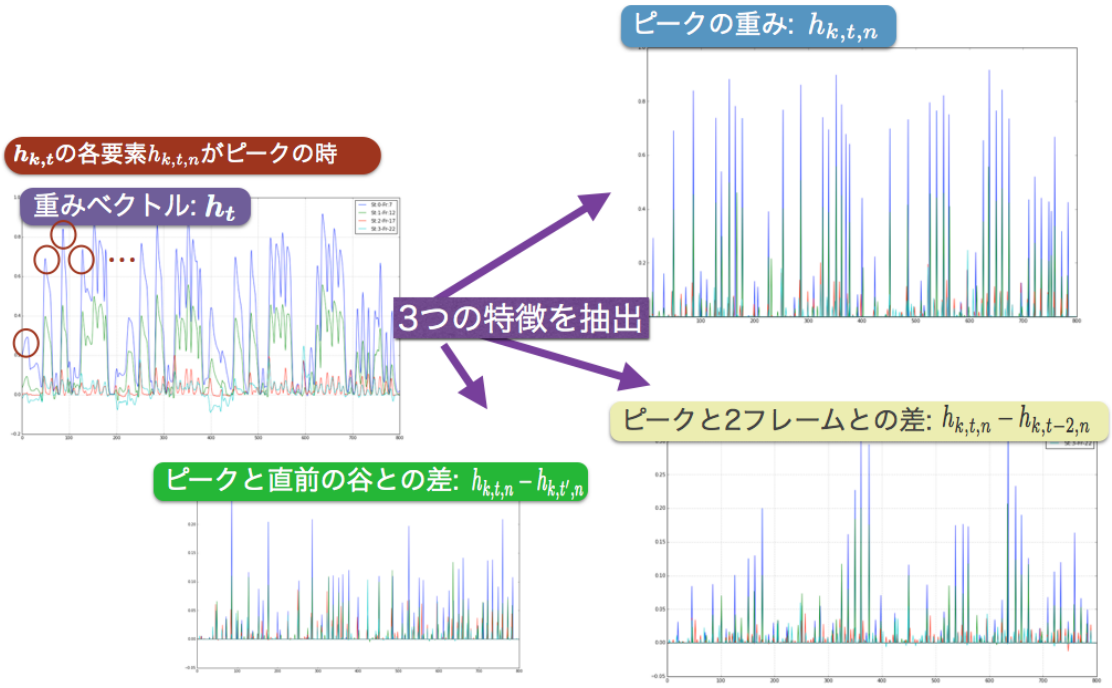

図 3.5: 学習に用いる入力ベクトル

3.3.2　本演奏に対する採譜及び、MIDI 形式への変換

上記のようにして弦ごとに入力ベクトルの準備を実行し、弦ごと各フレットごとの特徴ベクトル $\boldsymbol{x}_{k,t,n}$ を 10ms 間隔で求める。その後、第 2 ステップの学習後のニューラルネットワークに $\boldsymbol{x}_{k,t,n}$ を入力し、出力ノードの値である $y_{k,t,n}$ を計算する。本研究では、この $y_{k,t,n}$ を発音時刻であるかどうかの値を表す発音スコアと称する。そして、その発音スコア $y_{k,t,n}$ が閾値 y_0 を上回っているとき、言い換えれば $y_{k,t,n} > y_0$ を満たす値のときに、対応するフレット n に対応する MIDI ノートナンバーの MIDI ノート・オンメッセージを出力する。

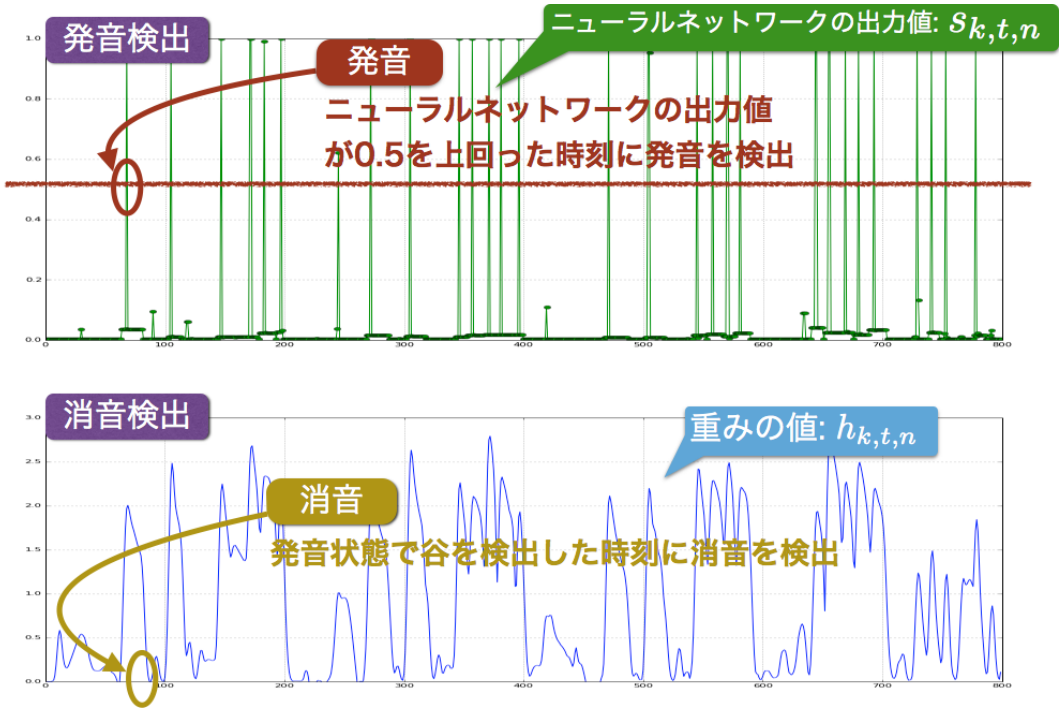

図 3.6: ニューラルネットワークを用いた発音・消音検出の例

3.4 評価実験

ニューラルネットワークによる発音検出手法の有効性を確認するために、単純な閾値処理による発音検出 (Baseline) とニューラルネットワークを用いて発音検出 (提案手法) を行った場合とで実験し、比較をする。

3.4.1 実験条件

各々手法において、同一演奏に対して Baseline 手法と提案手法を用いて採譜を行った。実験に際して、各々の手法に用いる閾値は、値による実験結果の変動を考慮し、各々の閾値とも計 10 個の値を用意し実験を行った。Baseline 手法の発音を検出するための閾値は 0.1 から 2.0 まで 0.2 刻みで変動させた。また、発音スコアの閾値は、0.0 から 1.0 の間で 0.1 刻みに閾値を変動させ採譜を行った。発音スコア

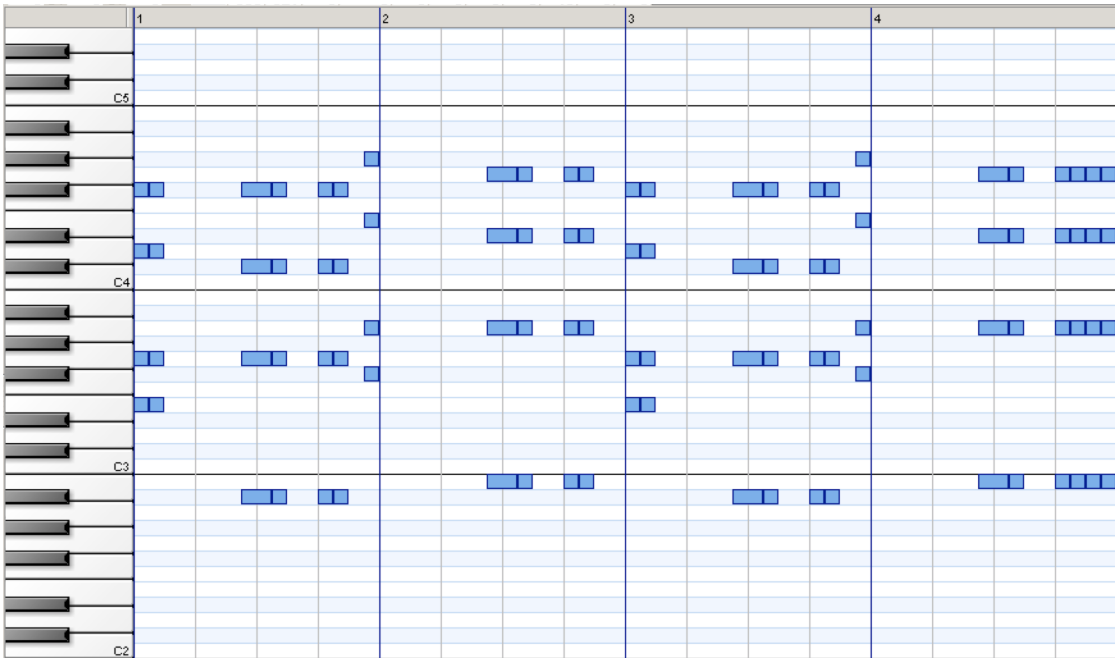

図 3.7: 実験に用いた第 2 予備演奏 1

の閾値とは、提案手法において発音時刻の判定に用いる閾値のことである。また、第 2 予備演奏を 2 つ (図 3.7, 図 3.8)(第 2 予備演奏 1, 第 2 予備演奏 2 とする) に設定した。これは文献 [17] から 2 つ選定した。第 2 予備演奏の選定基準は、16 ビートのカッティングフレーズであり、発音数が多いことである。この 2 つの第 2 予備演奏を用いて以下の学習データを作成し、ニューラルネットワークにより学習を行う:

1. 第 2 予備演奏 1 のみを用いた学習データ (学習データ 1)
2. 第 2 予備演奏 2 のみを用いた学習データ (学習データ 2))
3. 第 2 予備演奏 1 と第 2 予備演奏 2 の両方を用いた学習データ (学習データ 3)

採譜の対象となるフレーズは、文献 [17] から選んだ 4 小節のフレーズ (BPM120) 計 79 個である。なお、表 3.1 における「章」は、文献 [17] の章を表す。各々の手

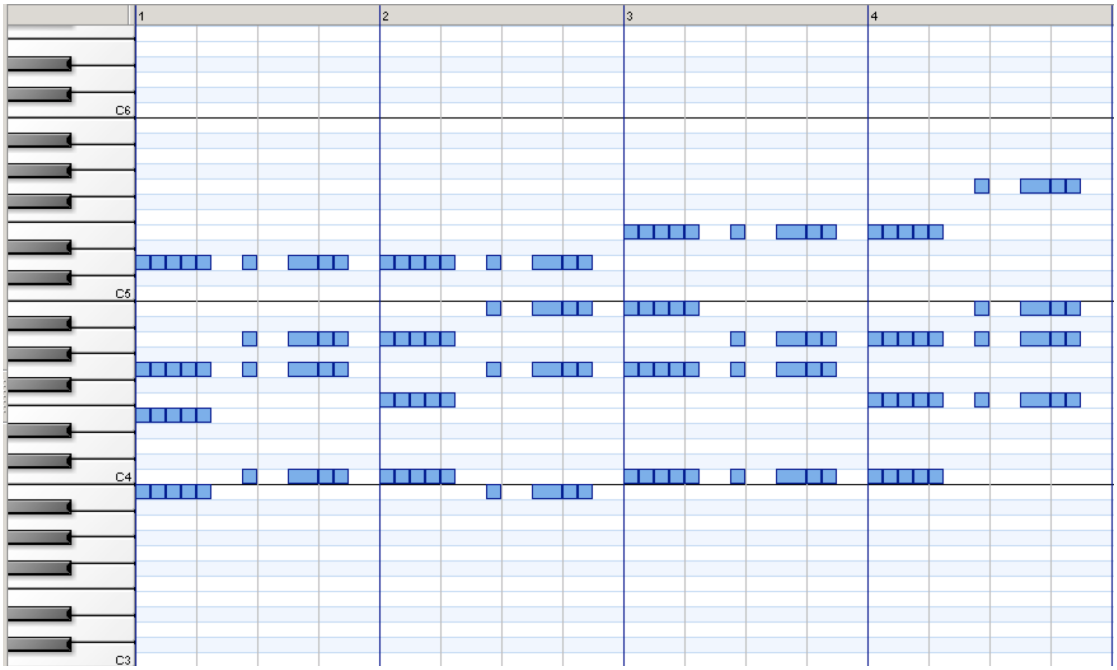

図 3.8: 実験に用いた第 2 予備演奏 2

法にて採譜対象を採譜し、音高と発音時刻を再現率、適合率と F 値を用いてそれぞれ評価し、章ごとに平均を取る。

3.4.2 実験結果、考察

Baseline 手法と提案手法の各々で最も高い F 値を記録した採譜結果の章ごとの平均値を表 3.1 に示す。表の可読性を考慮し、トラックごとの結果でなく、類似したトラックがまとまっている章ごとの平均値を結果として載せている。また、Baseline 手法により採譜した結果の閾値による変化の様子を図 3.9、及び提案手法での採譜結果の閾値による変化の様子を図 3.10 に示す。

まず、Baseline 手法と提案手法における採譜結果の章ごとの平均値について考察する。提案手法の再現率は、学習データ 1,2,3 全てにおいても平均 0.5 以上になり、最大の平均値は 0.559 という結果となった。特に、学習データ 3 においては、

0.6 以上の章が 5 つあり (1,2,5,6,13 章)、5 章では 0.838 という高い結果となった。Baseline 手法の平均 0.530 と比較すると、僅かではあるが、高い結果となった。一方、提案手法の適合率は、Baseline 手法より高い結果となった再現率とは逆に、学習データの順に平均 0.465, 0.463, 0.513 と Baseline 手法の平均 0.503 と比較すると学習データ 3 を除いて低い結果となった。そして、提案手法の F 値は、順に平均 0.488, 0.506, 0.526 となり、学習データ 3 のみが Baseline 手法の平均 0.516 を僅かながら上回る結果となった。

次に、各々の採譜手法における閾値の変化による採譜結果の変動を考察する。はじめに提案手法においては、学習データ 1 と 3 にて、再現率は閾値が 0.5 以下の区間において 0.5 を上回る結果となった。また、適合率は閾値がら 0.5 以下の区間では 0.5 を下回っているものの、0.5 以上では 0.5 を上回っている。学習データ 2 については、閾値が 0.3 において最大となる結果となった。なお学習データによる違いついては、学習データ 2 では 5,6 弦の発音がなかったために、学習済みのニューラルネットワークにおいて 5,6 弦の発音検出上手くできなく、他の学習データと比較して低い F 値となってしまったと考えられる。

一方 Baseline 手法では、閾値が 0.1 上がるにつれ、再現率は約 0.1 低下し、適合率は約 0.1 上昇するような結果となった。そのため、再現率と適合率ともに単調に変動している。

3.4.3　採譜結果例

Baseline 手法 (単純な閾値処理) による採譜結果と提案手法 (ニューラルネットワークを用いた発音検出処理) による採譜結果の一例を図 3.11、図 3.12、図 3.13 に示す。

平均的な結果として、Track 47-2 の採譜結果を図 3.11 に示す。Baseline 手法では、1 小節目などにおいて存在しない音が多く採譜されているが、提案手法では存在しない音が削減されている。その結果、適合率が Baseline 手法が 0.659 に対し、

表 3.1: Baseline 手法と提案手法の章単位での比較

	単純な閾値処理			提案手法								
				学習データ 1			学習データ 2			学習データ 3		
章	R	P	F	R	P	F	R	P	F	R	P	F
1 章	0.640	0.509	0.552	0.636	0.422	0.503	0.642	0.376	0.473	0.611	0.458	0.521
2 章	0.647	0.489	0.555	0.624	0.483	0.538	0.661	0.490	0.563	0.635	0.574	0.595
3 章	0.579	0.496	0.531	0.468	0.483	0.472	0.603	0.502	0.541	0.525	0.560	0.539
4 章	0.536	0.510	0.519	0.554	0.509	0.529	0.576	0.479	0.521	0.562	0.524	0.541
5 章	0.731	0.557	0.585	0.759	0.399	0.503	0.809	0.431	0.550	0.838	0.437	0.561
6 章	0.538	0.410	0.465	0.531	0.459	0.491	0.581	0.427	0.491	0.539	0.460	0.496
7 章	0.580	0.601	0.564	0.562	0.624	0.569	0.640	0.569	0.590	0.636	0.668	0.635
8 章	0.528	0.577	0.540	0.499	0.481	0.488	0.544	0.520	0.528	0.518	0.536	0.525
9 章	0.401	0.489	0.431	0.415	0.418	0.413	0.425	0.423	0.422	0.416	0.451	0.430
10 章	0.464	0.403	0.429	0.476	0.346	0.397	0.472	0.326	0.376	0.442	0.350	0.382
11 章	0.432	0.621	0.505	0.445	0.621	0.510	0.466	0.589	0.516	0.470	0.643	0.535
12 章	0.313	0.600	0.411	0.343	0.502	0.407	0.373	0.429	0.399	0.399	0.567	0.468
13 章	0.621	0.527	0.541	0.568	0.505	0.504	0.652	0.522	0.559	0.611	0.541	0.543
14 章	0.412	0.442	0.422	0.395	0.425	0.407	0.456	0.425	0.436	0.441	0.473	0.451
15 章	0.576	0.407	0.474	0.460	0.362	0.384	0.558	0.521	0.463	0.520	0.554	0.418
最終章	0.475	0.413	0.422	0.480	0.400	0.424	0.485	0.377	0.415	0.484	0.416	0.437
平均	0.530	0.503	0.516	0.513	0.465	0.488	0.559	0.463	0.506	0.540	0.513	0.526

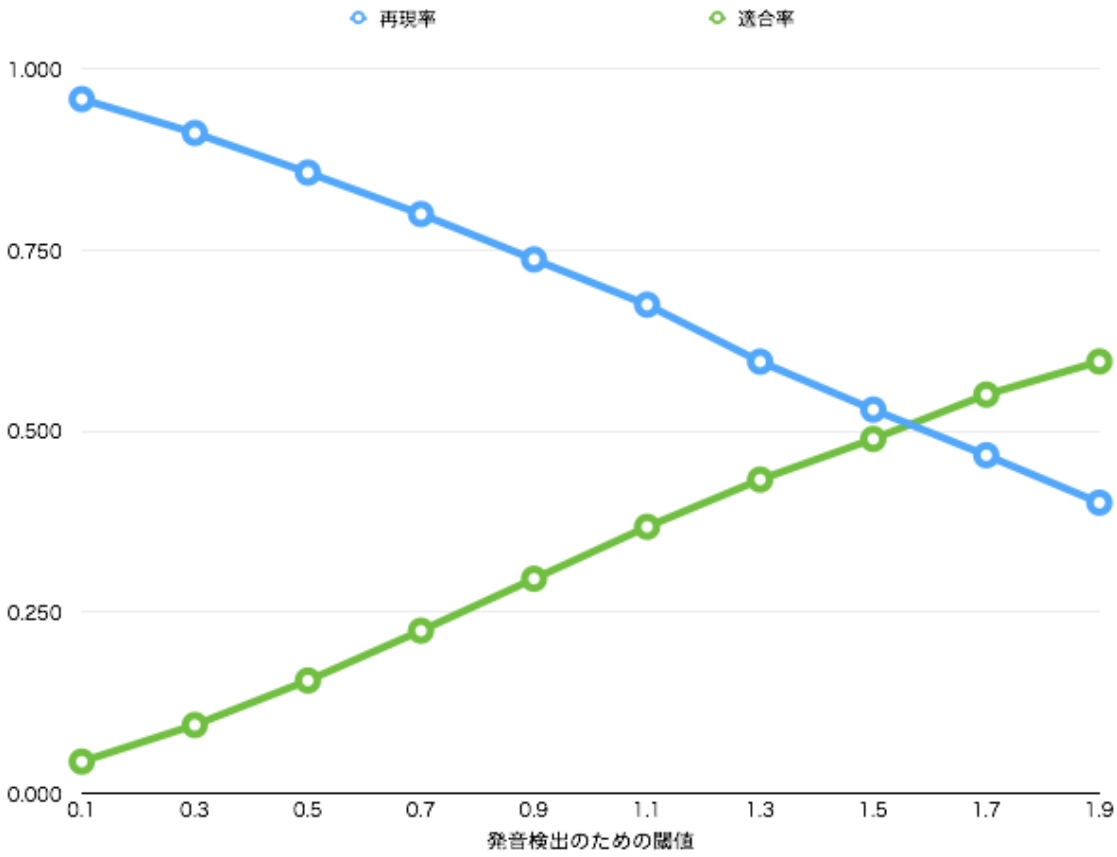

図 3.9: Baseline 手法 (単純な閾値処理) による採譜結果の再現率と適合率

提案手法では 0.692 となり、0.033 向上した。しかし、それに伴い存在する音も削減されてしまっており、再現率は Baseline 手法が 0.562 だが、提案手法では 0.556 と 0.006 低下した。再現率の低下より適合率の向上の方が大きかったため、F 値は Baseline 手法が 0.607 なのに対して、提案手法では 0.617 となった。

次に、精度が高かった結果として、Track 09-1 の採譜結果を図 3.12 に示す。両手法において、1,2,3,4 小節の 1 拍目において、連続した短い音が採譜されている。Baseline 手法では、フレーズ全体にわたり存在しない音の採譜が多くされているが、提案手法ではそれらの音が削減されている。その結果、適合率が Baseline 手法の 0.465 に比べ、提案手法では 0.661 となり、0.196 の向上となった。再現率は、両手法とも 0.72 と同じである。以上のことにより、F 値は、Baseline 手法の 0.565

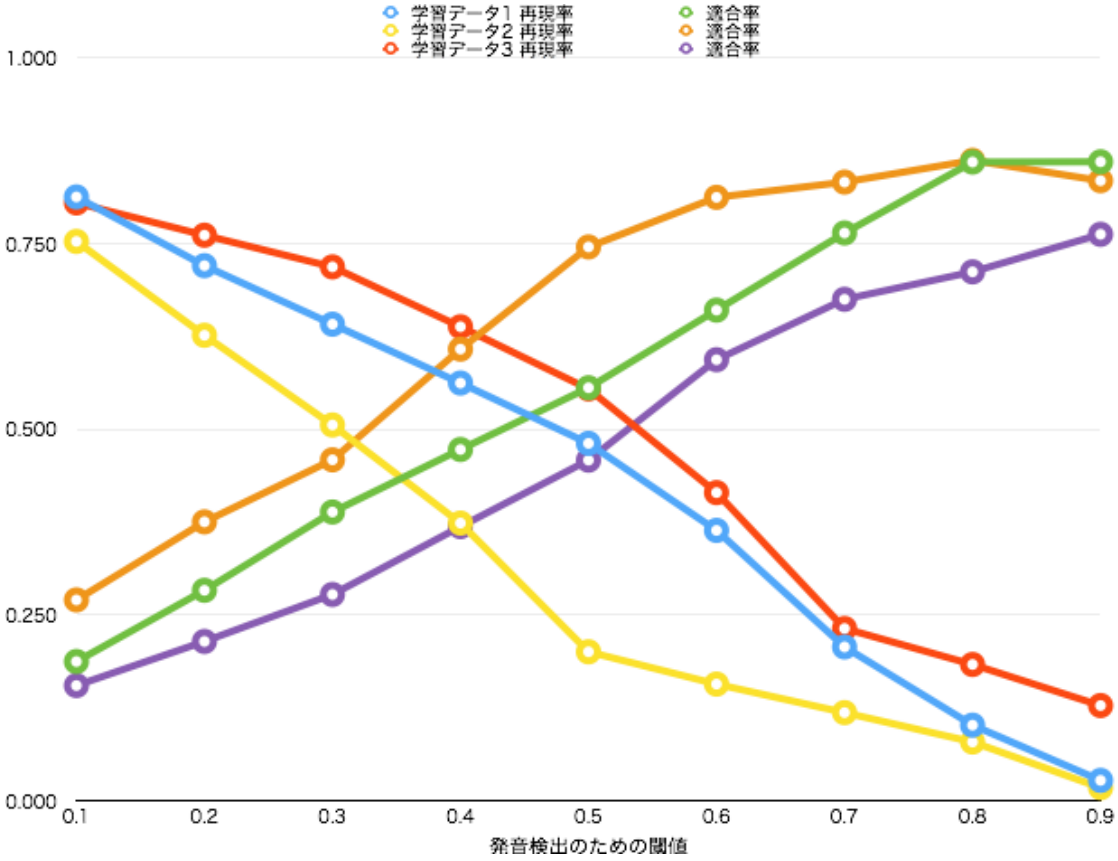

図 3.10: 提案手法による採譜結果の再現率と適合率

に対し、提案手法では 0.689 となり、0.124 の向上となった。

最後に、精度が低かった結果として、Track 70-2 の採譜結果を図 3.13 に示す。この採譜結果では、両手法においてフレーズ全体で存在する音が大きく欠落する結果となった。そのことが影響し、再現率は、Baseline 手法では 0.115、提案手法では 0.038 と著しく低い結果となった。また、Baseline 手法と提案手法ともに 2,4 小節目において倍音の採譜をしており、適合率は、Baseline 手法が 0.130、提案手法が 0.023 となった。以上の結果より、F 値は Baseline 手法が 0.122、提案手法が 0.029 と低い結果となった。これらの理由として、低音弦の基底推定が上手く行っておらず、低音弦の低フレットの音の欠落が多くなったためと考えられる。

3.5　おわりに

本章では、NMF を用いて音響信号処理による採譜手法について述べた。本章のまとめを以下に示す。

- 本来の NMF はオンラインに採譜することができない問題があった。そこで、予備演奏を演奏してもらうことにより、NMF をオンラインにて採譜できるように改良をした。また、倍音や演奏者のくせを考慮した発音検出ができるようにするために、更に予備演奏を追加し、ニューラルネットワークを用いた発音検出手法を提案した。

- 提案した手法の有効性を確認するために、79 フレーズの音高と発音時刻を対象に再現率、適合率、F 値を用いた評価実験を行った。提案手法において、閾値が真ん中である 0.5 の時に最大の F 値が観測される結果となり、閾値決定が容易にできる結果となった。更に、閾値を様々に変化させ、観測された各々の手法の最も良かった結果との比較においても、提案手法が Baseline 手法よりも再現率、再現率、適合率ともに 0.01 向上する結果となり、F 値の平均は 0.526 となった。

図 3.11: 平均的な採譜結果 (Track 47-2)

図 3.12: 精度が高かった採譜結果 (Track 09-1)

図 3.13: 精度が低かった採譜結果 (Track 70-2)

第4章　MIDIギターとNMFによる音響信号処理の統合による採譜手法

本章では、MIDI ギターと第 3 章で提案した NMF による音響信号処理の統合による採譜手法について述べる。まず、解決すべき MIDI ギターの問題点を提起する。そして、複数の統合方法について述べ、それらに対して評価実験を行う。

4.1　はじめに

MIDI ギターは、第 1 章でも述べたように、ギター演奏をリアルタイムで自動採譜することが可能であり、シンセサイザーや DTM の用途において有用である。しかし、MIDI ギターの採譜精度はジャンルや奏法において違いがある。特に、ファンクで頻繁に演奏されるゴーストノートを多用する 16 ビートのコードカッティングの演奏で 3 つの問題が頻繁に発生する。3 つの問題とは以下のものである:

1. 存在する音の欠落 (問題 1)
2. 存在しない音の採譜 (問題２)
3. 連続する短い音の融合 (問題 3)

MIDI ギターと NMF での音響信号処理の採譜結果では、誤り傾向に違いがある。MIDI ギターでは、ミュートした弦をピッキングした際などに発生するゴースト

ノートにより存在しない音が採譜される傾向にある。一方、NMF による音響信号処理による採譜では、倍音による採譜などにより存在しない音が採譜される傾向にある。また、MIDI ギターの方が連続した短い音がつながりやすい傾向にあったり、NMF の方が連続する短い音を採譜しやすいなど各々の採譜結果に異なる傾向が見られる (図 4.1)。

したがって、MIDI ギターと音響信号処理の出力結果の各々の傾向を考慮した統合により、MIDI ギターの 3 つの問題の音の補正、削減が課題となる。本研究では、音楽のジャンルをファンクに絞り、各々の傾向を利用し、統合によりこれらの問題を解決する。

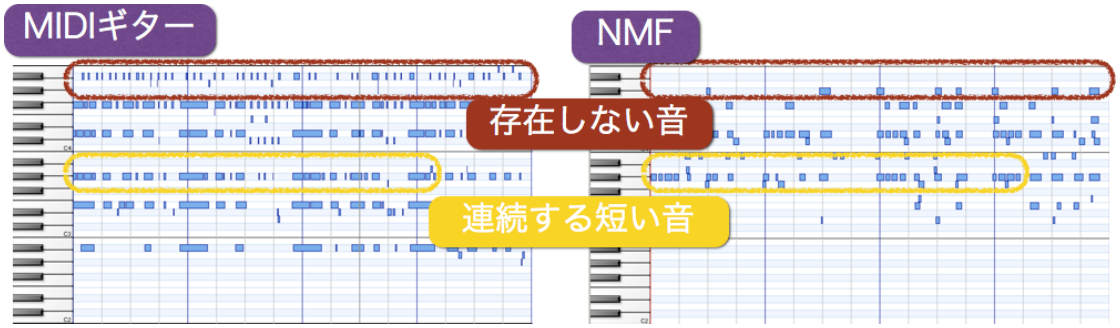

図 4.1: 採譜処理における誤り傾向の違いの一例

4.2　統合手法

本研究では、上記の 3 つの問題を解決するために、次の各手法を試すこととする:

1. 発音スコアの積による統合 (統合手法 1)
2. 発音スコアの和による統合 (統合手法 2)
3. ニューラルネットワークによる統合 (統合手法 3)

4.2.1 発音スコアの積による統合

MIDI ギター、音響信号処理ともに適合率の低さが問題であると考え、適合率向上のために、MIDI ギターと音響信号処理の両方で「発音している」と判断されたフレットのみ「発音している」とみなす方法である。弦 k ごとに、各時刻 t、フレット番号 n に対して 3.3 節の手法によりニューラルネットワークの出力値 $y_{k,n,t}$ を求める。音響信号処理と並行して MIDI ギターの出力を受け取り、

$$z_{k,n,t} = \begin{cases} 1-\alpha & (\text{時刻 } t \text{ において弦 } k \text{ の第 } n \text{ フレットが発音中}) \\ \alpha & (\text{上記以外}) \end{cases}$$

を求める。ここで、α は MIDI ギターの出力にない音も出力できるようにするパラメータであり、$0 \leq \alpha < 1$ である。これらの値の積 $w_{k,n,t} = y_{k,n,t} z_{k,n,t}$ を求め、弦 k ごとに時刻 t において $\max_n w_{k,n,t} > w_0$（w_0：あらかじめ決めた閾値）を満たすとき、$\hat{n} = \mathrm{argmax}_n\, w_{k,n,t}$ を求め、第 $\hat{n}$ フレットを発音したとみなして、対応するノートナンバーのノートオンメッセージを出力する。

4.2.2 発音スコアの和による統合

弦 k ごとに、各時刻 t、フレット番号 n に対して 3.3 節の手法により $y_{k,n,t}$ を求める。音響信号処理と並行して MIDI ギターの出力を受け取り、4.2.1 節と同様に $z_{k,n,t}$ を求める。これらの値の和 $w_{k,n,t} = y_{k,n,t} + z_{k,n,t}$ を求め、弦 k ごとに時刻 t において $\max_n w_{k,n,t} > w_0$（w_0：あらかじめ決めた閾値）を満たすとき、$\hat{n} = \mathrm{argmax}_n\, w_{k,n,t}$ を求め、第 $\hat{n}$ フレットを発音したとみなして、対応するノートナンバーのノートオンメッセージを出力する。

4.2.3　ニューラルネットワークに統合

ニューラルネットワークによる学習

音響信号処理による採譜の際に用いるニューラルネットワークの入力層に MIDI ギターによる出力を表すノードを追加する (図 4.2)。そして、4.2.1 節と同様に、弦 k ごとに、各時刻 t、フレット番号 n に対して 3.3 節の手法により $y_{k,n,t}$ を求める。音響信号処理と並行して MIDI ギターの出力を受け取り、

$$z_{k,n,t} = \begin{cases} 1-\alpha & (\text{時刻 } t \text{ において弦 } k \text{ の第 } n \text{ フレットが発音中}) \\ \alpha & (\text{上記以外}) \end{cases}$$

を求める。ここで、α は MIDI ギターの出力にない音も出力できるようにするパラメータであり、$0 \leq \alpha < 1$ である。そして、$h_{k,t,n}$ がピークの時に以下の処理を実行し、弦ごとかつフレットごとのデータセットを作成する:

1. $h_{k,t,n}$ から以下の特徴ベクトル $\boldsymbol{x}_{k,t,n}$ を抽出する:

$$\boldsymbol{x}_{k,t,n} = (h_{k,t,n},\ \ h_{k,t,n} - h_{k,t-2,n},\ \ h_{k,t,n} - h_{k,t',n},\ \ z_{t,n})$$

2. 入力ベクトル $\boldsymbol{x}_{k,t,n}$ の対となる出力値 $s_{k,t,n}$ は以下の様に定義される:

$$s_{k,t,n} = \begin{cases} 1\ (t \text{ がフレット } n \text{ に対して発音時刻の時}) \\ 0\ (t \text{ がフレット } n \text{ に対して発音時刻でない時}) \end{cases}$$

3. $\boldsymbol{x}_{k,t,n}$ と $s_{k,t,n}$ をデータ・セットにしてニューラルネットワークの学習を行う。

3.3 節と同様にして、学習の収束を考え、データ・セットにおける $s_{k,t,n}$ の 1 と 0 の数のバランス比を同じにする。なお、演奏時の僅かなずれを考慮し、時刻 t における $\boldsymbol{x}_{k,t,n}$ は、予め設定した正解の発音時刻の前後 6 フレーム以内であれば時刻 t を発音時刻とみなす。

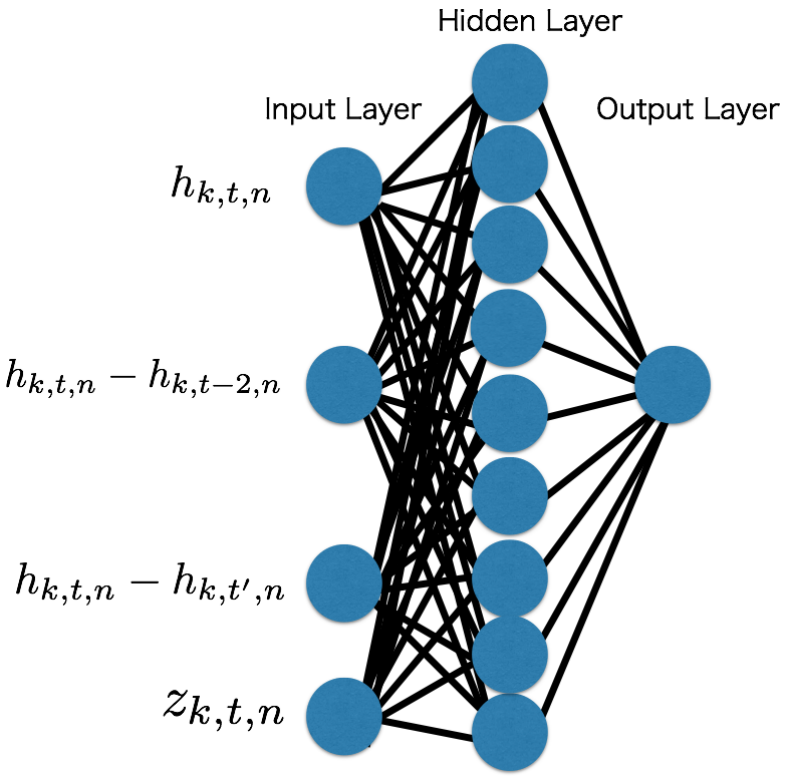

図 4.2: MIDI ギターと音響信号処理の統合に用いるニューラルネットワーク

MIDI への変換

学習のときと同様にして入力ベクトル $\boldsymbol{x}_{k,t,n}$ を作成し、学習済みのニューラルネットワークに入力する。そして、ニューラルネットワークの出力値がある閾値を越えた時に発音とみなし、対応するノートナンバーのノートオンメッセージを出力する。

4.3　評価実験

4.3.1　実験手法

上記の 3 手法において、発音スコアの閾値を 0.0 から 1.0 の範囲で 0.1 刻みに変動させて評価実験を行った。採譜対象の演奏は、3 章と同じく文献 [17] から選んだ 79 フレーズ (BPM120) である。なお、表 4.1, 表 4.2, 表 4.3 における「章」は、文献 [17] の章を表す。採譜対象全 79 トラックを各々の手法にて採譜をし、音高と発音時刻を再現率、適合率、F 値を用いて評価し、章ごとの平均を結果とした。なお、MIDI ギターの再現率、適合率、F 値は定数であるので、閾値が変化しても値は変化しない。また、統合手法 3 以外では、ニューラルネットワークは 3 章で用いたものの内最も F 値が高かった学習データ 3 のものを使用した。

4.3.2　実験結果、考察

統合手法 1: 発音スコアの積による統合

発音スコアの積による統合手法の採譜結果の中で最も高い F 値を記録した採譜結果の章ごとの平均値を表 4.1 に示す。表の可読性と紙面のスペースを考慮し、トラックごとの結果でなく、類似したトラックがまとまっている章ごとの平均値を結果として載せている。また、発音スコアの積による統合手法で採譜した結果の閾値による変化の様子を図 4.3 に示す。

はじめに、表 4.1 の章ごとに採譜結果をまとめたものについて考察をする。まず全体的な傾向から考察すると、再現率は、MIDI ギターの平均 0.528 と比較して、統合手法では平均 0.595 と 0.067 精度向上させることができた。MIDI ギターの採譜結果では、連続した短い音が融合して 1 つの長い音として採譜されていた。しかし、音響信号処理の採譜結果と統合したことにより、音響信号処理の採譜の連続する短い音を採譜しやすい傾向が功を奏して、音が分割され採譜されるようになっ

たと考えられる。その結果、連続した短い音の融合が削減され、再現率が 0.067 向上した。一方、適合率は、提案手法は平均 0.660 となり、MIDI ギターの平均 0.258 より 0.402 も精度が向上した。積を取り統合したことにより、存在しない音が削減され、適合率の向上になったと考えられる。再現率と適合率の向上から F 値は 0.626 となり MIDI ギターの 0.347 より 0.313 も精度向上した。よって、再現率、適合率、F 値ともに MIDI ギターよりも採譜精度を向上させることができた。

次に、章ごとに採譜結果を観察すると、再現率は、計 11 章 (1,2,3,4,6,7,8,10,12,13, 最終章) にて精度を向上させることができた。しかし、計 5 章 (5,9,11,14,15 章) にて採譜精度が低下する結果となってしまった。5 章においては、単音カッティングと呼ばれる奏法を用いたフレーズの章であり、音響信号処理の結果が単音の採譜精度が低い結果であったことから再現率の低下が起こってしまったと考えられる。第 9 章においても、低音弦を利用した単音カッティングのフレーズが含まれていることから再現率の低下が起こったものと考えられる。第 11 章においては、様々な形のボイシングでコードを演奏するフレーズの章であることから、音高の移り変わりが多い。そのような傾向かつ短い音の場合において、MIDI ギターの採譜結果と音響信号処理による採譜結果に時刻のずれがある時などに、その時刻において MIDI ギターと音響信号処理とで共通の発音がない結果となってしまった。その結果、発音スコアが低下し、発音が検出されなかったために、再現率が低下してしまったと考えられる。14 章においては、1 小節単位の所々で単音カッティングのフレーズがあることから再現率の低下が起きたと考えられる。15 章でも、ミュートした弦での単音カッティングのフレーズが 1 フレーズあり、再現率の低下が起きたと推測される。以上のことから、統合手法 1 では主に、単音カッティングの奏法において再現率の低下が引き起こされている。

適合率については、計 16 章である全ての章において適合率の向上という結果となった。特に、5,11,13 章においては、適合率が 0.5 以上も向上している。しかし、5,11,13 章は、再現率の低下が見られた章でもあり、MIDI ギターと音響信号処理

とで共通の発音が観測されていないため、再現率の低下し、適合率の向上が起きたと考えられる。

F 値は、適合率が大きく向上したことが影響し、全ての章において F 値の向上という結果となった。特に、5 章においては、MIDI ギターの 0.195 よりも 0.472 向上した 0.667 という結果となった。

次に 、表 4.3 の閾値を変化させた時の採譜結果の変動について考察する。再現率は閾値が 0.8 以上の場合を除き、統合手法は MIDI ギターの 0.528 を上回ることができた。更に、0.2 から 0.6 の区間においては、適合率も 0.5 を上回っており、それに伴い F 値も 0.5 以上の結果となっている。幅広い閾値で高い F 値を観測できたことから閾値設定に依存しないことが言える。したがって、統合により MIDI ギターの問題点である、存在する音の欠落と連続する短い音の融合の補正ができたことにより精度が向上できた言える。閾値 0.8 以上において、MIDI ギターの結果を下回ってしまった理由としては、あまりに発音スコアの閾値を高く設定しすぎたためと考えられる。

最後に、それぞれの結果をまとめると、再現率は MIDI ギターよりも平均 0.1、最大 0.4 向上させることができた。また、適合率は MIDI ギターよりも平均 0.25、最大 0.5 向上させることができた。そのことにより、F 値は、MIDI ギターと比較して平均 0.1 以上向上させ、最大 0.3 向上させることができた。 以上のことから、積による統合により、存在する音の欠落、存在しない音の採譜、連続する短い音の融合の 3 つの問題を軽減でき、採譜精度を向上させたことが示された。

統合手法 2: 発音スコアの和による統合

発音スコアの和による統合手法の採譜結果の中で最も高い F 値を記録した採譜結果の章ごとの平均値を表 4.2 に示す。統合手法 1 の発音スコアの積による統合と同様に、表の可読性と紙面のスペースを考慮し、トラックごとの結果でなく、類似したトラックがまとまっている章ごとの平均値を結果として載せている。また、

発音スコアの和による統合手法で採譜した結果の閾値による変化の様子を図 4.4 に示す。

はじめに、表 4.2 の章ごとに採譜結果をまとめたものについて考察をする。統合手法 1 と類似した結果となった。まず全体的に結果は、再現率は、MIDI ギターの平均 0.528 と比較して、統合手法では平均 0.612 と 0.084 と精度向上させることができた。積による統合の平均 0.595 と和による統合の平均 0.612 と比較すると、0.17 向上している。また、適合率は、提案手法は平均 0.642 となり、MIDI ギターの平均 0.258 より 0.384 も精度が向上した。再現率と適合率が向上した理由は、統合手法 1 の発音スコアの和による統合と同様と考えられる。再現率と適合率が共に向上したことにより、F 値は 0.626 となり MIDI ギターの 0.347 より 0.28 も精度向上した。したがって、再現率、適合率、F 値ともに MIDI ギターよりも採譜精度を向上させることができた。

次に、章ごとの採譜結果を観察する。再現率は、計 12 章 (1,2,3,4,6,7,8,10,11,12,13, 最終章) にて精度を向上させることができた。統合手法 1 の積による統合と比較すると、11 章において MIDI ギターの平均 0.425 より 0.32 向上させ平均 0.457 とした。したがって、再現率が向上した合計が 1 章多い結果となった。しかし、計 4 章 (5,9,14,15 章) にて採譜精度が低下する結果となってしまった。5 章においては、単音カッティングと呼ばれる奏法を用いたフレーズの章であり、音響信号処理の結果が単音の採譜精度が低い結果であったことから再現率の低下が起こってしまったと考えられる。低下の原因は、統合手法 1 と同様と考えられる。しかし、11 章においては統合手法 1 との差異がある。差異の理由としては、11 章が様々な形のボイシングでコードを演奏するフレーズの章であることから、音高の移り変わりが多い。統合手法 1 では、そのような傾向かつ短い音の場合において、MIDI ギターの採譜結果と音響信号処理による採譜結果に時刻のずれがある時などに、その時刻において MIDI ギターと音響信号処理とで共通の発音がない結果となってしまった。そして、積を取った結果、発音スコアが低下し、発音が検出されなかったた

めに、再現率が低下してしまったと考えられる。一方、和による統合では、低い値同士でも足しあわされるため、積による統合よりも発音スコアが高くなり、発音を多く検出できたために再現率が向上したと考えられる。

適合率については、計 16 章である全ての章において適合率の向上という結果となった。特に、統合手法 1 と同様に、5,11,13 章においては、適合率が約 0.4 以上も向上している。しかし、5,13 章は、再現率の低下が見られた章でもある。MIDI ギターと音響信号処理とで共通の発音が観測されていないため、再現率の低下は起きたが、適合率の向上が起きたと考えられる。

F 値は、適合率が大きく向上したことが影響し、全ての章において F 値の向上という結果となった。特に、5 章では、MIDI ギターの 0.195 よりも 0.472 向上した 0.667 という結果となった。これは、統合手法 1 の積による統合と類似した結果となった。

次に、表 4.4 の変動における採譜結果の変動について考察する。再現率は閾値が 0.8 以上の場合を除き、統合手法は MIDI ギターの 0.528 を上回ることができた。更に、0.4 から 0.6 の区間においては、適合率も 0.5 を上回っており、それに伴い F 値も 0.5 以上の結果となっている。幅広い閾値で高い F 値を観測できたことから閾値設定に依存しないことが言える。したがって、統合により MIDI ギターの問題点である、存在する音の欠落と連続する短い音の融合の補正ができたことにより精度が向上できた言える。閾値 0.8 以上において、MIDI ギターの結果を下回ったのは、統合手法 1 と同様の理由と考えられる。このように、統合手法 1 の積による統合と類似した結果となった。

最後に、それぞれの結果をまとめると、再現率は MIDI ギターよりも平均 0.1、最大 0.4 向上させることができた。また、適合率は MIDI ギターよりも平均 0.25、最大 0.5 向上させることができた。そのことにより、F 値は、MIDI ギターと比較して平均 0.1 以上向上させ、最大 0.3 向上させることができた。 以上のことから、和による統合により、存在する音の欠落、存在しない音の採譜、連続する短い音

の融合の 3 つの問題を軽減でき、採譜精度を向上させたことが示された。

ニューラルネットワークによる統合

ニューラルネットワークによる統合手法の採譜結果の中で最も高い F 値を記録した採譜結果の章ごとの平均値を表 4.3 に示す。上記した 2 つの統合手法と同様に、表の可読性と紙面のスペースを考慮し、トラックごとの結果でなく、類似したトラックがまとまっている章ごとの平均値を結果として載せている。また、ニューラルネットワークによる統合手法で採譜した結果の閾値による変化の様子を図 4.5 に示す。

はじめに、表の 4.3 の章ごとに採譜結果を観察すると、再現率は、全ての章において精度を向上させることができた。しかし、適合率は、計 2 章 (5,7 章) を除き、低下する結果となってしまった。また、適合率は 0.20 以下の結果が 9 個以上となった。この理由としては、音の欠落を補うのに伴い存在しない音の採譜が増えてしまったためと考えられる。以上の結果から、F 値は計 6 章 (3,4,5,7,8,13 章) では 0.03 前後と僅かながらに向上したものの、その他では F 値が低下する結果となった。

続いて、表 4.5 の変動における採譜結果の変動について考察する。再現率は全ての場合において、統合手法は MIDI ギターの 0.528 を上回ることができた。更に、0.1 から 0.8 の区間においては、0.750 以上の結果となった。しかし、全ての閾値において適合率は 0.3 以下と低い結果となっている。これに伴い F 値も 0.4 以下の結果となっている。したがって、統合により音の欠落を補正できるようにはなったものの、それにともなって存在しない音までもを採譜するようになってしまっていると考えられる。閾値が 0.9 のときに最も良い結果が得られた理由としても、閾値が高くなったことにより存在しない音が削減されたためと考えられる。

最後に、それぞれの結果をまとめると、再現率は MIDI ギターよりも平均 0.2、最大 0.46 向上させることができた。しかし、適合率は MIDI ギターよりも平均 0.06、

最大 0.28 低下した。そのことにより、F 値は、MIDI ギターと比較して平均 0.02 低下し、最大 0.23 低下する結果となった。 以上のことから、ニューラルネットーワークによる統合にでは、存在する音の欠落、連続する短い音の融合の 2 つの問題を僅かながらに軽減できたが、存在しない音の出力を助長する結果となり、採譜精度が低下した。

まとめ

発音スコアの積による統合と発音スコアの和による統合において、MIDI ギターのよりも F 値を最大 0.3 近く向上させることができた。発音スコアの積による統合と和による統合において各々の結果は類似したものとなったが、再現率は発音スコアの和の方が高い結果となり、適合率は発音スコアの積の方が高い結果となった。F 値は、積と和による統合ともに同じ結果となった。しかし、ニューラルネットワークによる採譜では一部の場合を除き採譜精度を低下した。

4.3.3　採譜結果例

MIDI ギターによる採譜結果と統合手法において最も精度が高かった発音スコアの積による統合の採譜結果の一例を図 4.6、図 4.7、図 4.8 に示す。

平均的な結果として、Track 47-2 の採譜結果を図 4.6 に示す。MIDI ギターでは、フレーズ全体にわたり存在しない音が多く採譜されているが、統合による採譜では存在しない音が削減されている。その結果、適合率が MIDI ギターが 0.48 に対し、統合による採譜では 0.83 となり、0.35 と大きく向上した。しかし、それに伴い存在する音も削減されてしまっておる。だが、再現率は MIDI ギターは 0.60 だが、統合による採譜では 0.67 と 0.07 向上した。再現率向上の理由として、、1 小節目の 1 拍目などで連続する短い音が融合していたものが、統合により連続した

音として採譜されるようになったためと考えられる。以上のことにより、F 値は MIDI ギターが 0.53 なのに対して、統合による採譜では 0.74 となった。

次に、精度が高かった結果として、Track 09-2 の採譜結果を図 4.7 に示す。MIDI ギターにおいて、1,2,3,4 小節の 1 拍目において、連続した短い音が融合して採譜されている。一方、統合による採譜では、それらの音が分割され、連続する短い音として採譜されている。このことにより、再現率は MIDI ギターが 0.61 に対し、統合による採譜は 0.80 と 0.19 向上した。また、MIDI ギターでは、1 小節目後半から 4 小節の最後にわたり存在しない音の採譜が多くされているが、統合による採譜ではそれらの音が削減されている。その結果、適合率が MIDI ギターの 0.25 に比べ、統合による採譜では 0.78 となり、0.53 の向上となった。そのような結果から、F 値は、MIDI ギターの 0.36 に対し、統合による採譜では 0.79 となり、0.43 と大幅な向上となった。

最後に、精度が低かった結果として、Track 69-2 の採譜結果を図 4.8 に示す。MIDI ギターの採譜結果は、存在しない音を 1,4 小節目などで多く採譜しているものの、存在する音を多く採譜しているため、再現率は 0.78、適合率は 0.28 となっている。一方、統合による採譜では、存在する音の欠落がフレーズ全体にわたって発生しており、再現率は 0.26、適合率は 0.41 となった。この理由として、NMF による音響信号処理において、低音弦の基底推定が上手く行っておらず、低音弦の低フレットの音の採譜ができずに存在する音の欠落が多くなり、その結果、積による統合の際に発音スコアが低く観測され発音が検出されなかったためと考えられる。

4.4 おわりに

本章では、MIDI ギターと NMF を用いた音響信号処理を統合による採譜手法について述べた。本章のまとめを以下に示す。

- MIDI ギターと NMF の音響信号処理をオンラインで統合する手法を 3 種類

提案した。1 つ目は発音スコアの積を用いた統合、2 つ目は発音スコアの和を用いた統合である。最後に、3 つ目としてニューラルネットワークを用いた統合である。

- 提案した 3 つの手法の有効性を確認するために、79 フレーズの音高と発音時刻を対象に再現率、適合率、F 値を用いた評価実験を行った。結果、発音スコアの積を用いた統合と発音スコアの和を用いた統合の 2 つの提案手法が MIDI ギターよりも再現率を、適合率、F 値とも向上させることができた。発音スコアの積においては、再現率は 0.067、再現率は 0.402、F 値 0.279 とそれぞれ向上し、F 値の平均は 0.626 となった。また、発音スコアの和による統合では、再現率は 0.084、適合率は 0.384、F 値は 0.279 とそれぞれ向上し、F 値の平均は 0.626 となった。しかし、ニューラルネットワークを用いた統合においては、再現率は 0.190 向上したものの、適合率は 0.066、F 値は 0.044 低下し、F 値の平均は 0.303 となった。

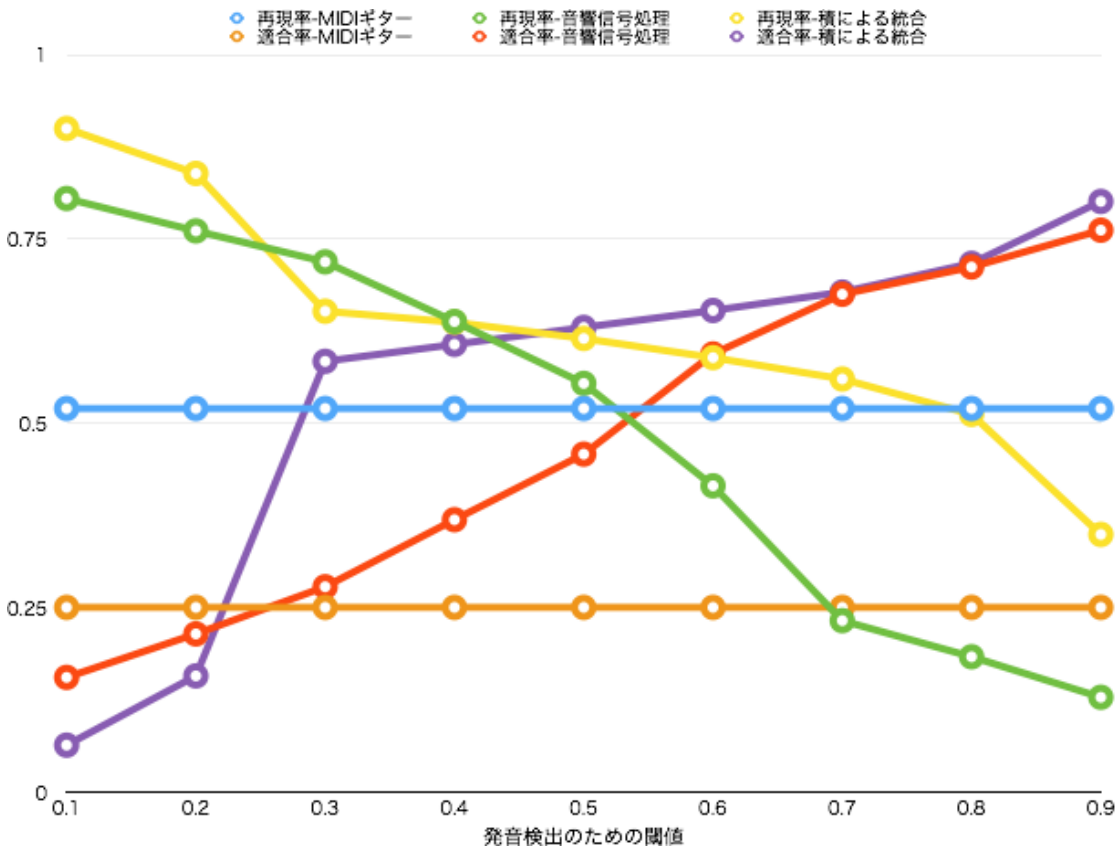

図 4.3: 発音スコアの積による統合を用いた採譜結果の再現率と適合率

表 4.1: 発音スコアの積による統合の採譜結果の再現率、適合率、F 値

	MIDI ギター			NMF			発音スコアの積による統合		
章	R	P	F	R	P	F	R	P	F
1 章	0.668	0.445	0.513	0.611	0.458	0.521	0.758	0.589	0.660
2 章	0.630	0.230	0.338	0.635	0.574	0.595	0.717	0.648	0.679
3 章	0.380	0.310	0.327	0.525	0.560	0.539	0.763	0.613	0.679
4 章	0.505	0.233	0.313	0.562	0.524	0.541	0.508	0.670	0.575
5 章	0.805	0.110	0.195	0.838	0.437	0.561	0.731	0.643	0.675
6 章	0.583	0.187	0.287	0.539	0.460	0.496	0.649	0.660	0.641
7 章	0.493	0.205	0.278	0.636	0.668	0.635	0.597	0.660	0.606
8 章	0.503	0.215	0.295	0.518	0.536	0.525	0.593	0.644	0.602
9 章	0.533	0.345	0.408	0.416	0.451	0.430	0.398	0.603	0.469
10 章	0.463	0.190	0.267	0.442	0.350	0.382	0.732	0.545	0.623
11 章	0.425	0.345	0.375	0.470	0.643	0.535	0.378	0.825	0.493
12 章	0.310	0.260	0.285	0.399	0.567	0.468	0.488	0.708	0.574
13 章	0.438	0.183	0.250	0.611	0.541	0.543	0.640	0.674	0.644
14 章	0.555	0.280	0.360	0.441	0.473	0.451	0.475	0.776	0.550
15 章	0.652	0.354	0.434	0.520	0.554	0.418	0.532	0.695	0.517
最終章	0.505	0.238	0.313	0.484	0.416	0.437	0.555	0.607	0.542
平均	0.528	0.258	0.347	0.540	0.513	0.526	0.595	0.660	0.626

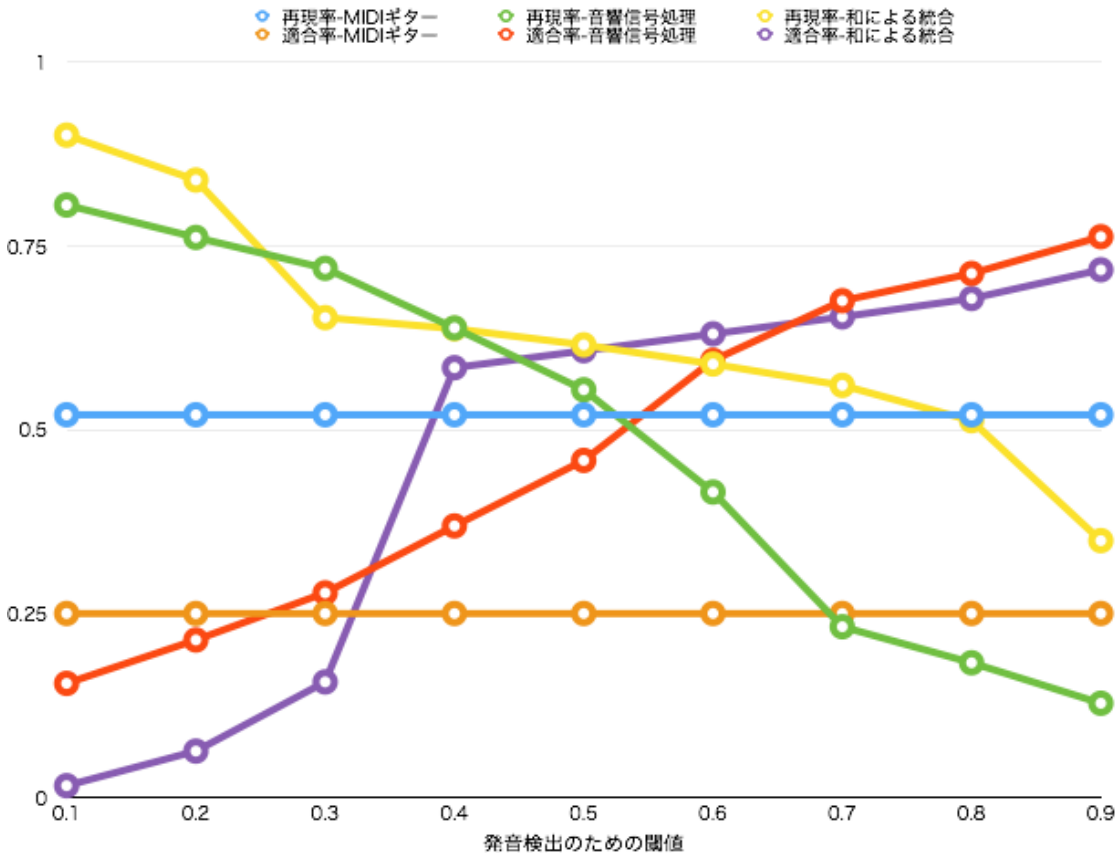

図 4.4: 発音スコアの和による統合を用いた採譜結果の再現率と適合率

表 4.2: 発音スコアの和による統合の採譜結果の再現率、適合率、F 値

	MIDI ギター			NMF			発音スコアの和による統合		
章	R	P	F	R	P	F	R	P	F
1 章	0.668	0.445	0.513	0.611	0.458	0.521	0.734	0.584	0.647
2 章	0.630	0.230	0.338	0.635	0.574	0.595	0.731	0.633	0.676
3 章	0.380	0.310	0.327	0.525	0.560	0.539	0.800	0.605	0.686
4 章	0.505	0.233	0.313	0.562	0.524	0.541	0.538	0.659	0.590
5 章	0.805	0.110	0.195	0.838	0.437	0.561	0.722	0.636	0.667
6 章	0.583	0.187	0.287	0.539	0.460	0.496	0.669	0.654	0.654
7 章	0.493	0.205	0.278	0.636	0.668	0.635	0.663	0.635	0.637
8 章	0.503	0.215	0.295	0.518	0.536	0.525	0.601	0.650	0.609
9 章	0.533	0.345	0.408	0.416	0.451	0.430	0.444	0.650	0.513
10 章	0.463	0.190	0.267	0.442	0.350	0.382	0.680	0.572	0.619
11 章	0.425	0.345	0.375	0.470	0.643	0.535	0.457	0.780	0.548
12 章	0.310	0.260	0.285	0.399	0.567	0.468	0.488	0.615	0.541
13 章	0.438	0.183	0.250	0.611	0.541	0.543	0.648	0.660	0.642
14 章	0.555	0.280	0.360	0.441	0.473	0.451	0.515	0.683	0.556
15 章	0.652	0.354	0.434	0.520	0.554	0.418	0.544	0.653	0.519
最終章	0.505	0.238	0.313	0.484	0.416	0.437	0.552	0.600	0.532
平均	0.528	0.258	0.347	0.540	0.513	0.526	0.612	0.642	0.626

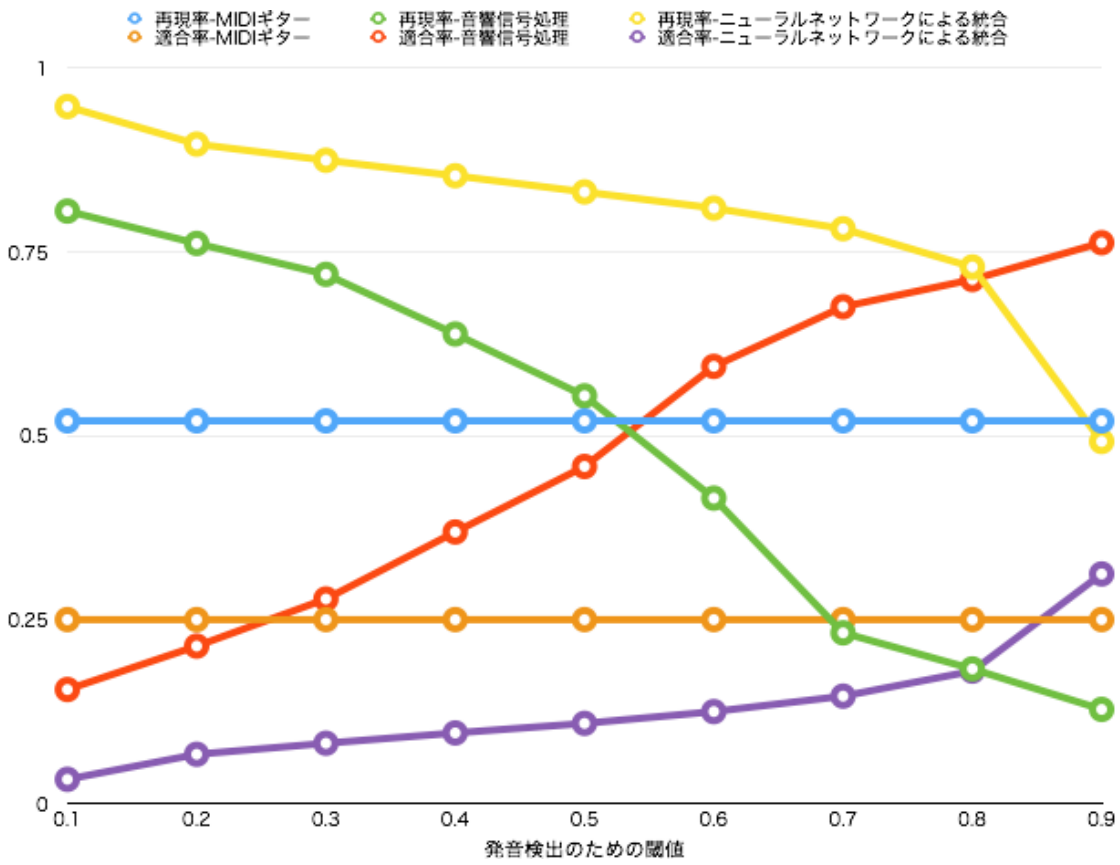

図 4.5: ニューラルネットワークによる統合を用いた採譜結果の再現率と適合率

表 4.3: ニューラルネットワークによる統合の採譜結果の再現率、適合率、F 値

	MIDI ギター			NMF			ニューラルネットワークによる統合		
章	R	P	F	R	P	F	R	P	F
1 章	0.668	0.445	0.513	0.610	0.460	0.520	0.907	0.166	0.279
2 章	0.630	0.230	0.338	0.640	0.570	0.600	0.788	0.172	0.281
3 章	0.380	0.310	0.327	0.530	0.560	0.540	0.836	0.228	0.357
4 章	0.505	0.233	0.313	0.560	0.520	0.540	0.766	0.204	0.321
5 章	0.805	0.110	0.195	0.840	0.440	0.560	0.857	0.140	0.234
6 章	0.583	0.187	0.287	0.540	0.460	0.500	0.732	0.152	0.251
7 章	0.493	0.205	0.278	0.640	0.670	0.640	0.730	0.233	0.345
8 章	0.503	0.215	0.295	0.520	0.540	0.530	0.692	0.203	0.313
9 章	0.533	0.345	0.408	0.420	0.450	0.430	0.528	0.162	0.248
10 章	0.463	0.190	0.267	0.440	0.350	0.380	0.695	0.134	0.222
11 章	0.425	0.345	0.375	0.470	0.640	0.540	0.631	0.224	0.328
12 章	0.310	0.260	0.285	0.400	0.570	0.470	0.505	0.192	0.278
13 章	0.438	0.183	0.250	0.610	0.540	0.540	0.780	0.189	0.297
14 章	0.555	0.280	0.360	0.440	0.470	0.450	0.637	0.198	0.301
15 章	0.652	0.354	0.434	0.520	0.550	0.420	0.713	0.329	0.298
最終章	0.505	0.238	0.313	0.480	0.420	0.440	0.685	0.150	0.240
平均	0.528	0.258	0.347	0.541	0.513	0.527	0.718	0.192	0.303

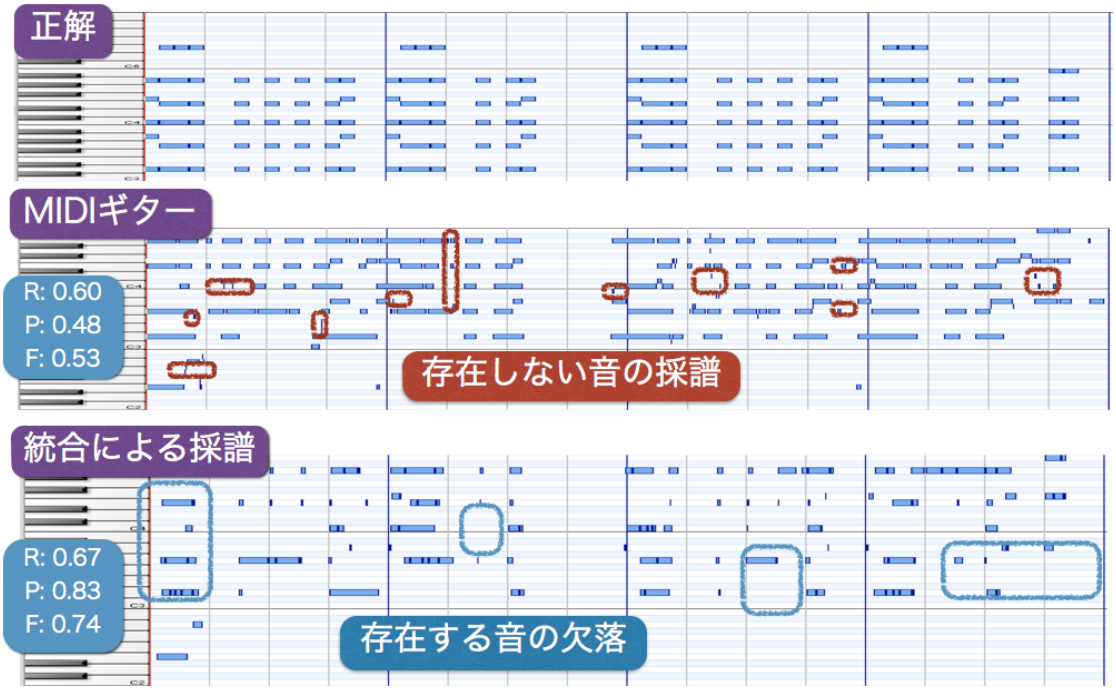

図 4.6: 平均的な採譜結果 (Track 47-2)

図 4.7: 精度が高かった採譜結果 (Track 09-2)

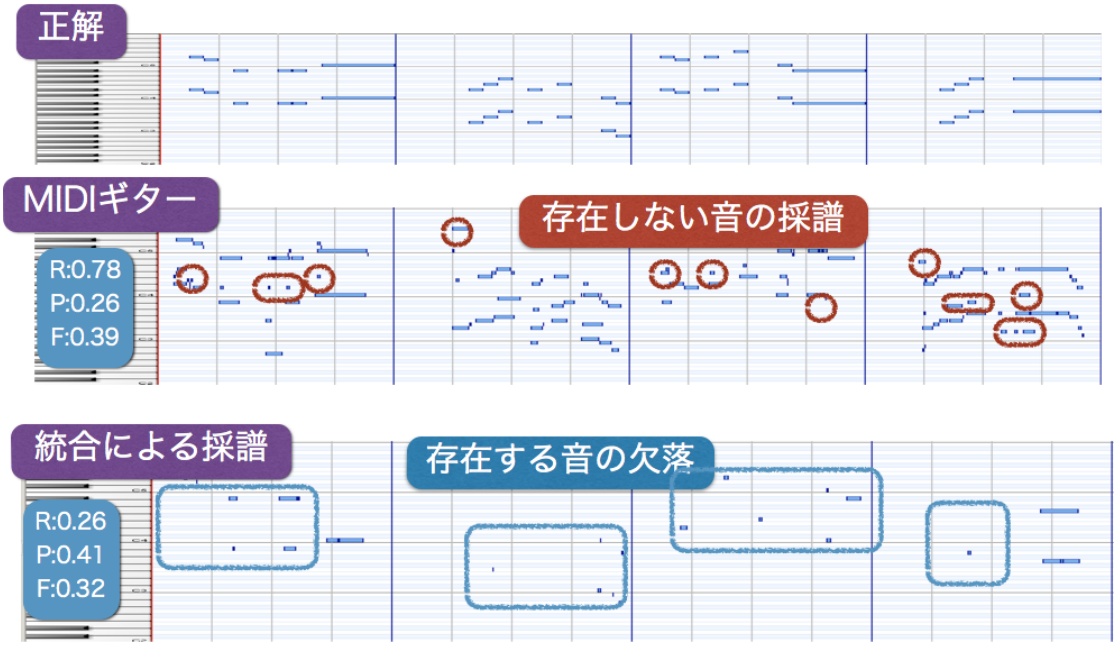

図 4.8: 精度が低かった採譜結果 (Track 69-2)

第5章　今後の課題

本章では、4章までで提案した手法についての今後の課題について述べる。

5.1　音響信号処理に用いるNMFの変更

本研究では、音響信号処理に採譜に基本的なNMF[12]を用いた。しかし、基底行列の推定に基本的なNMFを用いたために、現状では低音源の低フレットの音の基底分解が上手く行えていないケースが存在する。そこで、音響信号処理に適するように改良されたNMF[18]が研究されている。この改良されたNMFは、音の立ち上がりや減衰などの音楽音響信号の時間に伴い変化するスペクトルに対応するように基底の分解を改良したものである。このNMFを用いることにより、基底行列の取得がより精度の高いものになり採譜精度の向上が図れる可能性がある。音響信号音響信号に用いるNMFの変更も今後の課題である。

5.2　第2予備演奏の選定

本研究では、NMFによる音響信号処理の発音検出のために第2予備演奏を用いた。ファンクの演奏におけるMIDIギターの採譜精度補正を目指して、第2予備演奏の選定方法は16ビートのカッティングフレーズであり、かつギターの全弦を使用していることとした。しかし、様々な予備演奏による採譜結果の違いについては未検証であるため、ユーザーの負担が小さく、採譜精度が高くなる予備演奏の制定については今後の課題である。

5.3 ニューラルネットワークでの学習に用いる特徴量の改良

NMF による音響信号処理においてニューラルネットワークによる閾値の調整手法を提案した。提案手法では、できるだけシンプルなネットワークを用いて学習を行った。しかし、学習に用いる入力ベクトルの妥当性については未検証である。したがって、採譜精度が効果的に向上できる特徴のベクトルの検証は今後の課題である。

5.4 身体的制約や画像処理の導入

本研究では、手の大きさなどの身体的な制約を導入したり、画像などの他モダリティを併用するような既存研究で提案されている手法を用いない範囲での採譜精度の向上を目指した。しかし、それらの他モダリティを用いるアプローチは本研究にも導入可能であり、導入することにより更なる精度向上も可能である。本研究の実用化も向上も考慮に入れ、これらのアプローチの導入も今後の課題である。

5.5 リアルタイム実装

本稿では、提案した統合手法はオンラインアルゴリズムであるため、リアルタイムで動作するように実装することは可能である。リアルタイムでの実装に際して、MIDI ギターと NMF による音響信号処理の各々の遅延など考慮が必要な事項があるが、MIDI ギターと NMF による音響信号処理をリアルタイムに行い、遅延時間や採譜結果の評価実験を行いたい。

第6章　結　　論

本研究では、MIDI ギターと NMF を用いた音響信号処理の統合によるギター演奏を対象とした自動採譜手法を提案した。MIDI ギターには、存在する音の欠落、存在しない音の採譜、連続する短い音の融合の 3 つの問題があった。統合によりこれらの問題を解決し、MIDI ギターの採譜結果よりも再現率、適合率、F 値を向上させることができた。

各章で述べた内容の要旨は次の通りである。

第 1 章では、本研究の背景と目的を明らかにした。自動採譜が様々な範囲に応用できること、特にギター演奏を対象とした自動採譜の需要が高いことを述べ、MIDI ギターと NMF による音響信号処理をオンラインで統合することにより目的の達成を目指すことを述べた。

第 2 章では、自動採譜の研究について述べ、特にギター演奏に対する自動採譜の研究にどのようなアプローチが存在するかを概観した。そして、本研究がギター演奏を対象とした自動採譜の研究の中でもどのような位置づけになるかを述べた。

第 3 章では、NMF を用いた音響信号処理による採譜手法について述べた。NMF をオンラインで採譜可能なように改良をし、更にニューラルネットワークを用いて発音検出を行う手法を提案した。各々の手法について評価実験を行い、提案手法において、閾値が真ん中である 0.5 の時に最大の F 値が観測される結果となり、閾値決定が容易にできる結果となった。更に、閾値を様々に変化させ、観測された各々の手法の最も良かった結果との比較においても、提案手法が Baseline 手法よりも再現率、再現率、適合率ともに 0.01 向上する結果となり、F 値の平均は 0.56

となった。

第4章では、MIDIギターとNMFを用いた音響信号処理による採譜処理の統合による採譜手法について述べた。発音は発音スコアの積による統合、発音のスコアの和による統合、ニューラルネットワークによる統合の3つの統合手法を提案した。評価実験の結果、発音スコアの積を用いた統合と発音スコアの和を用いた統合の2つの提案手法がMIDIギターよりも再現率を、適合率、F値とも向上させることができた。発音スコアの積においては、再現率は0.067、再現率は0.402、F値0.279とそれぞれ向上し、F値の平均は0.626となった。また、発音スコアの和による統合では、再現率は0.084、適合率は0.384、F値は0.279とそれぞれ向上し、F値の平均は0.626となった。しかし、ニューラルネットワークを用いた統合においては、再現率は0.19向上したものの、適合率は0.066、F値は0.044低下し、F値の平均は0.303となった。

第5章では、本研究に残された課題と今後の展望について述べた。

最後に、本研究の貢献をまとめる。MIDIギターとNMFを用いた音響信号処理の統合によるギター演奏の自動採譜を提案した。身体的制約や画像処理などのトップダウンな情報や他モダリティを用いない範囲での採譜精度の高精度化する手法を提案した。他モダリティを用いない範囲での高精度化のため、今後は違う採譜手法を用いての統合などの参考になることが期待される。また、本研究の応用に関しては、DTMにおけるMIDI入力は当然のことながら、ジャムセッションシステムの開発などにも応用可能である。本研究が、自動採譜、更に言えば音楽情報処理の発展に少しでも貢献できたならば幸いである。

参考文献

[1] Musical Instrument Digital Interface: http://www.midi.org/

[2] 有元慶太ら他: “楽器固有の高調波構造モデルを用いたギター演奏に対する多重音高推定手法”, 日本音響学会論文集, pp. 585–586, 2006.

[3] M. Goto: “A Real-Time Music Scene Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals”, *Speech Communication*, Vol. 43, No. 4, pp.311-329, 2004.

[4] K. Yazawa, K. Itoyama and H. G. Okuno: “Automatic Transcription of Guitar Tablature From Audio Signals In Accordance With Player’s Proficiency”, *ICASSP*, pp.3146–3150, 2014.

[5] A. M. Barbancho and A. Klapuri: “Automatic Transcription of Guitar Chords and Fingering From Audio”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol.20, pp.915–921, 2012.

[6] M. Paleari, B. Huet, A. Schutz, and D. Slock, “A multimodal approach to music transcription”, *ICIP*,pp. 93–96, Oct. 2008.

[7] T. Yamagami and K. Itou: “A Bimodal Music Dictation Method for Composition Support by using Guitar Performance Video”, *Proceeding of IPSJ National convention 2014*, pp.2-365–2-366, 2014.

[8] X. Fiss and A. Kwasinski: “Automatic Real-Time Electric Guitar Audio Transcription”, *ICASSP*, pp.373-376, 2011.

[9] P. D. O’Grady and S. T. Rickard: “Automatic Hexaphonic Guitar Transcription Using Non-Negative Constraints”, *ISSC*, 2009.

[10] J. Hartquist : “Real-time Musical Analysis of Polyphonic Guitar Audio”, *Master Thesis, The Faculty of California Polytechnic State University*, 2012.

[11] 内山裕貴ら他: “調波音・打楽器音分離手法を用いたギター・ベースギターの自動採譜”, 情報処理学会第 *76* 回全国大会, pp. 363–364, 2014.

[12] D.D. Lee and H.S. Seung: “Learning The Parts of Objects with Nonnegative Matrix Factorization”, *Nature*, vol.401, pp.788-791, 1999.

[13] 青木直史ら他: “画像処理によるギター運指動作のキャプチャリング”, 電子情報通信学会総合大会, D-11-110, 2005.

[14] You Rock Guitar: http://yourockguitar.com/

[15] YAMAHA EZ-EG: http://www.yamaha.co.jp/design/products/2000/ez-eg/

[16] A. B. Israel and T.N.E. Greville: “Generalized Inverses: Theory and Applications”, *New-York, Springer*, 2003.

[17] 山口和也: “16 ビートが身につく! ファンクで覚える大人のカッティング”, リットー・ミュージック, 2013.

[18] 中野允裕他: “可変基底 NMF に基づく音楽音響信号の解析”, 情報処理学会研究報告, 2010.

謝　　辞

本論文を作成するにあたり、北原鉄朗准教授から、丁寧かつ熱心なご指導を賜りました。また、北原研究室のメンバーとは苦楽を共にし、時には切磋琢磨し、成長を共にしました。特に、同学年である小暮計貴氏とは、お互いが研究が行き詰まった時には議論をし、様々な助言をいただきました。また、本論文の審査に当たり、副査を担当してくださった斎藤明教授と尾崎知伸准教授には、本質的な鋭いご指摘やご意見を頂きました。他にも様々な人々の助けがあり、研究を進めることができました。

本論文をまとめることができたのは、多くの方々のご尽力、ご支援のおかげであります。ここに心からの感謝の意を表します。