学習するジャムセッションシステム：演奏者固有のフレーズの獲得

浜中雅俊[†]　後藤真孝[†††,‡]　麻生英樹[‡]　大津展之[†,‡,‡‡]

[†]筑波大学　[†††]科学技術振興事業団さきがけ研究 21「情報と知」領域

[‡]産業技術総合研究所　[‡‡]東京大学

{m.hamanaka, m.goto, h.asoh, otsu.n}@aist.go.jp

本稿では, 実在する人間の演奏者固有のフレーズを模倣した仮想演奏者と人間の演奏者がインタラクションできるようなジャムセッションシステムについて述べる．演奏者固有のフレーズを模倣するためには，ジャムセッションの演奏記録から，フレーズを再利用可能な形で切り出し，フレーズデータベース上に蓄積する必要がある．本研究では，自動でフレーズデータベースを作成するための手法として，(1）確率モデルに基づいた発音時刻のクォンタイズとそのモデルパラメータの教師なし推定法，(2）ボロノイ線図を用いたグルーピング手法によるフレーズ分割，(3）局所自己相関関数を用いたフレーズからの特徴抽出法とフレーズの空間配置法，を提案する．

A Learning Session System: Acquisition of Player's Musical Phrases

Masatoshi Hamanaka[†]　Masataka Goto[†††,‡]　Hideki Asoh[‡]　Nobuyuki Otsu[†,‡,‡‡]

[†]Univ. of Tsukuba　[†††]" Imformation and Human Activity" PRESTO, JST

[‡]National Institute of Advanced Industrial Science and Technolory (AIST)　[‡‡]The Univ. of Tokyo

Mbox 0604. 1-1-1Umezono, Tsukuba, Ibaraki 305-8568 Japan

This paper describes a jam session system that enables a human player to interplay with virtual players, each of which imitates musical phrases of a human player. In order to imitate musical phrases of a human player, it is necessary to build a database of phrases extracted from session recording. We propose the following three methods for creating a database automatically: (1) a probabilistic-model-based quantization method for estimating the positions of onset times in a score and an unsupervised estimating method for the model parameters, (2) a phrase dividing method using the Voronoi diagram, and (3) a method of extracting the phrase characteristics by using autocorrelations and a method of locating phrases in a space.

1.はじめに

本研究の目的は，ある演奏者を模倣した個性をもつ仮想演奏者と人間の演奏者とが一緒にセッションすることができるシステムを実現することである．演奏者の個性には様々な側面があるが，たとえばジャムセッションでは，相手の演奏に対してどのような演奏で反応するかに演奏者の個性が表れる．そのような個性をここでは，演奏者の振る舞いと呼ぶ．また，ジャムセッションではフレーズや音自体にも個性が表れる．あたかも実在する人間の演奏者のような仮想演奏者を生成するためには，これらの個性を何らかの形でモデル化・獲得する必要がある．このような個性の獲得が可能となれば，親しい演奏者や，自分よりも演奏能力の高い演奏者，既に亡くなった演奏者を模倣した個性をもつ仮想演奏者といつでもインタラクションすることができるし，自分自身のモデルを用いた仮想演奏者とジャムセッションを行うことも可能となる．

従来のジャムセッションシステム[1][2]では，人間の演奏に追従した演奏を仮想演奏者にさせることに主眼がおかれていたため，仮想演奏者に個性を持たせるには至らなかった．一方，文献[3][4]では，個性データベースと呼ばれるルール群を導入することにより，個性の違いを設定することを可能としていた．文献[5][6]では，システム外部から変更可能なパラメータを複数用意することにより，各演奏者が主導権を握る程度を様々に変化させることを可能としていた．しかし，これら[3]-[6]は，パラメータやルールの調整を行うこと

により，異なる振る舞いの仮想演奏者を設定することはできても，実在する演奏者の振る舞いを模倣するようなモデルを獲得することは困難であった．

一方，文献[7][8]では，あらかじめ用意した長さ 8 小節の入出力演奏パターン 30 対をニューラルネットで学習することにより，人間の 8 小節の演奏に対して，8 小節の演奏パターンを出力するシステムを実現した．この研究は，演奏者のモデルを学習により獲得しようとした点が優れており，複数の演奏パターンの補間により新たな演奏パターンが生成できることを示していた．しかし，実際のセッションの演奏では，相手の同じような演奏に対してまったく同じ演奏で反応する可能性は低く，同じ入力に対して複数の出力が考えられるため，文献[7][8]の手法を用いて，演奏記録から演奏者のモデルを直接学習することは困難であった．

これに対して我々はこれまで，入力演奏を一段抽象化した印象空間と出力演奏を一段抽象化した意図空間を構成することにより，模倣しようとする演奏者がどんな演奏に対してどのような即興演奏を行ってきたかを，印象空間から意図空間への関数として，演奏記録から統計的に学習する手法を提案した[9]．そして，実際に 12 コーラスのセッションの演奏記録から演奏者の振る舞いのモデルが獲得できることを示した．得られたモデルをもった仮想演奏者 2 人と人間の演奏者でセッションした結果，仮想演奏者は人間の演奏者と対等な立場で，ソロや伴奏を交代しながら，模倣した人間の演奏者に近い振る舞いをすることが確認できた．

文献[9]では，仮想演奏者は事前に人間の演奏者から学習した演奏者の振る舞いのモデルに基づき演奏意図を決定し，フレーズデータベースの中からその意図に対応したフレーズを次々と接続することで演奏を生成していた．フレーズデータベースには 1 小節から 8 小節の長さの演奏が収められていたが，それらは，手作業で作成していたため，仮想演奏者は演奏者固有のフレーズまでは模倣できなかった．

そのような模倣を可能とするためには，模倣したい演奏者の即興演奏から演奏者固有のフレーズを切り出して，データベースを自動生成しフレーズを再利用する必要がある．そこで我々はこれまで，要素技術としてクォンタイズ[10]とフレーズ分割[11]を研究してきた．文献[10]では，フレーズ切り出しの際ゆらぎを除去せずに行うと，フレーズの接続時につなぎ目が不自然となる問題を解決するため，確率モデルを用いて，伴奏に合わせて弾いた演奏の発音時刻から，元々演奏者が弾こうとした正規化された楽譜上の発音時刻を推定するクォンタイズ手法を提案した．また，正解データが人手でラベル付けしてある演奏記録を用いてモデルパラメータを教師つき学習した結果，市販のシーケンスソフトウェアに搭載されているクォンタイズを超える性能をもつことを示した．しかし，文献[10]の手法では，モデルパラメータを教師つき学習していたため，新たな演奏者の演奏記録をクォンタイズすることは容易でなかった．一方，文献[11]では，ボロノイ線図を用いてピアノロール譜面上の音符のグルーピングが可能であることを示し，ボロノイ線図によるグルーピングが，GTTM 理論[12]を用いて人間が行ったグルーピングと近い結果が得られることを確認した．しかし，この手法が実際にフレーズ分割に有効であるかは確認されていなかった．

本研究では，ジャムセッションの演奏からフレーズを自動で切り出し，データベースに蓄積する手法を構築し，切り出されたフレーズを再利用することにより，演奏者固有のフレーズまでも模倣した仮想演奏者を実現する手法について述べる．以下，2 節では，ジャムセッションシステムの全体像について述べる．次に 3 節では，文献[10]のクォンタイズ手法を発展させ，モデルパラメータの教師なし学習手法を提案し，正解データの与えられない条件でもパラメータ推定が可能であることを示し，実際にフレーズ切り出しの前処理としてクォンタイズを行う．4 節では，文献[11]のグルーピング手法を用いて，実際にジャムセッションの演奏からフレーズ切り出しが可能であることを示す．5 節では，切り出したフレーズをジャムセッションシステムで再利用する手法について述べ，最後に 6 節でまとめと今後の課題を述べる．

2. セッションシステムの概要

本ジャムセッションシステムは，人間と仮想演奏者合わせて 3 人のギタートリオが，お互いに主従関係を持たずにソロや伴奏を自由に繰り返しながら 12 小節 1 コーラスで典型的なブルース進行の曲を演奏するものである．調は A，テンポ一定，コード進行は固定である．システムの入出力には MIDI を用い，人間の演奏者は MIDI ギターを演奏する．仮想演奏者は，その内部に**印象空間**，**意図空間**，**演奏者の振る舞いのモデル**をもつ（**図 1**）．印象空間は入力演奏を抽象化した空間で，1 人の演奏者の演奏は印象空間上の印象ベクトルであらわされる．意図空間は仮想演奏者が演奏を生成するときに使用する演奏パターンを主観的類似度に基づき配置した空間で，1 つの演奏パターンは，意図空間上の意図ベクトルで表される．演奏者の振る舞いのモデルは，模倣しようとする演奏者が参加したセッションの演奏記録を用いて，印象ベクトルから意図ベク

図 1：ジャムセッションシステム

トルへの関数を学習したものである．仮想演奏者は，3人の演奏の過去 12 小節の 3 つの印象ベクトルから, 演奏者の振る舞いのモデルに基づき次の意図ベクトルを決定する．そして，その意図ベクトルに対応したフレーズをデータベースから選択し出力する．

本稿では，ジャムセッションの演奏記録からフレーズデータベースを自動生成し，データベースに蓄積されたフレーズを再利用して仮想演奏者の演奏を生成する手法について述べる．フレーズデータベースを自動で生成するためには，以下の課題がある．

(1) 発音時刻ゆらぎの除去

フレーズ切り出しの前処理として，発音時刻ゆらぎを除去する必要がある．もし，ゆらぎを除去せずにフレーズ切り出しすると，生成時にフレーズを接続する際に，つなぎ目が不自然となる部分が生じてしまう．たとえば，はしったリズムのフレーズと，もたったリズムのフレーズが接続された部分では，拍の間隔が不規則となり聞き苦しい演奏となってしまう．

(2) フレーズの切り出し

演奏記録からフレーズを切り出すためには，音楽の階層的な構造を考慮して，その切れ目となる部分で分割する必要がある．もし，小節線の位置などで機械的に演奏記録を分割してしまうと，階層構造を壊してしまう場合があり，そのようなフレーズの断片を再利用することは困難である．

(3) フレーズの意図空間への配置

切り出したフレーズを再利用して仮想演奏者が演奏を生成するためには，仮想演奏者がどのような意図のときにそのフレーズを使用すれば良いか分からなければならない．すなわち，各フレーズに対応する意図ベクトルを求め，意図空間上に配置する必要がある．

上記の課題の解決法を以下(2.1〜2.3 節)に述べる．

2.1 確率モデルに基づくクォンタイズ

発音時刻を確率モデルにより推定しクォンタイズすることで，発音時刻のゆらぎを除去する．ジャムセッションのように, 8 分 3 連音符や 16 分音符が頻繁に入れ替わるような演奏を正しくクォンタイズするには，確率モデルによる発音時刻の推定が有効であることがわかっている[10]．しかし，文献[10]では，教師データが与えられ，モデルパラメータが既知という条件で，クォンタイズの性能を評価していたため，新たな演奏者の演奏に対して適切なモデルパラメータを用いてクォンタイズすることはできなかった．

そこで本研究では，**Baum-Welch アルゴリズム**を用いて教師データが与えられない条件でモデルパラメータを推定する手法を提案する．その際，学習データの不足によりモデルパラメータが正しく求まらない問題を解決するため，**ヘルドアウト補間法**を導入する．

2.2 ボロノイ線図を用いたフレーズの分割

ボロノイ線図によるグルーピング手法を用いて，ジャムセッションの演奏をフレーズ分割する．音楽の階層的な構造を考慮しながらグルーピングするための音楽理論としては，GTTM の**グルーピング構造分析**がある．しかし，GTTM のグルーピング構造分析では単旋律の楽曲が対象で，和声に関する分析が欠如しているため，ジャムセッションの演奏のフレーズ分割に直接用いることはできない．また，ルールが定量的に定義されておらず，複数のルール間の順位が不明確で競合が起きるなど，理論を計算機上へ実装することは困難だと考えられている．

そこで本研究では，GTTM によるグルーピング結果と近い結果を出すボロノイ線図を用いたグルーピング手法[11]によりグルーピングを行い, その結果を用いて, フレーズ分割を行うことを提案する

2.3 正準相関分析による意図ベクトルの算出

意図空間とフレーズの物理的特徴との相関を正準相関分析により求め，その相関を用いて意図ベクトルを計算する．意図空間は，フレーズを主観的類似度に基づき配置した空間である．主観的類似度は，フレーズ間の類似度を被験者実験によって求めたものである．

本研究では，主観的に類似したフレーズが意図空間上で近くに配置されるように，多次元尺度法の 1 つである MDA-OR を用いて，フレーズを配置する．そして，意図空間と出力演奏との相関は，**局所自己相関関数**によって得られたフレーズの物理的特徴と意図ベクトルとを**正準相関分析**することにより求める．

3.教師なしクォンタイズ

本節では，ジャムセッションの発音時刻を，楽譜上の正規化された位置へとクォンタイズする手法を説明し，教師データが与えられない条件で確率モデルのモデルパラメータを推定する手法を提案する．

演奏者が同じ演奏を繰り返し弾いた場合でも, MIDI のレベルでまったく同じ演奏となることは稀であり，演奏動作の微妙な差や，演奏の表情づけの差などにより発音時刻がゆらぐ．これを，元々演奏者が弾こうとした発音時刻（小節・拍内の正規化された位置）系列から，ゆらぎのある実際に演奏された発音時刻系列を求める順問題とすると，演奏された発音時刻系列から演奏者が元々弾こうとした正規化された発音時刻系列を求めるクォンタイズは, その逆問題 となる（**図 2**）．ここでは，前者のゆらぎの生じる過程を「**発音時刻ゆらぎの順モデル**」としてモデル化する．そして，この順モデルから求めた「**発音時刻ゆらぎの逆モデル**」を用いて逆問題を解き，クォンタイズを実現する．

3.1 発音時刻の遷移とゆらぎのモデル

発音時刻ゆらぎの順モデルは，元々弾こうとした正

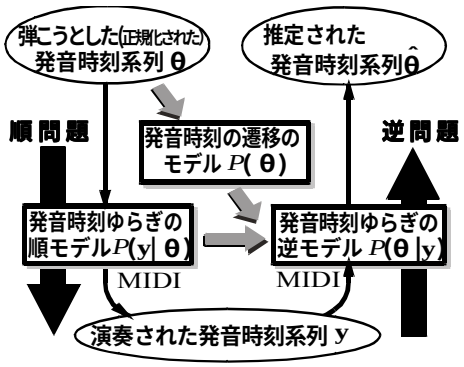


図 2：クォンタイズにおける順モデルと逆モデル

規化された発音時刻系列を θ ，演奏された発音時刻系列を y としたとき，$P(y|\theta)$ で表される．このとき，逆モデルは，Bayes の定理より式(1)で表される．

$$P(\theta|y)=\frac{P(y|\theta)P(\theta)}{P(y)} \qquad (1)$$

$P(\theta)$は，θ に対する事前分布であり，演奏者がどのような発音時刻系列を弾きやすいかを表す．逆問題の解 $\hat{\theta}$ は，式(1)を最大化する θ である（式(2)）．

$$\hat{\theta}=\arg\max_{\theta} P(\theta|y)=\arg\max_{\theta} P(y|\theta)P(\theta) \qquad (2)$$

本研究では，発音時刻ゆらぎの順モデル $P(y|\theta)$ と発音時刻の遷移のモデル $P(\theta)$ を組み合わせたモデルを隠れマルコフモデル（HMM）で定式化する．HMM はマルコフ的な隠れ状態遷移モデル $P(\theta)$ と各状態における出力確率分布 $P(y|\theta)$ を組み合わせたモデルである．我々が観測することができるのは，出力確率分布から得られる出力のみで，どの状態にいるかを観測することはできない．

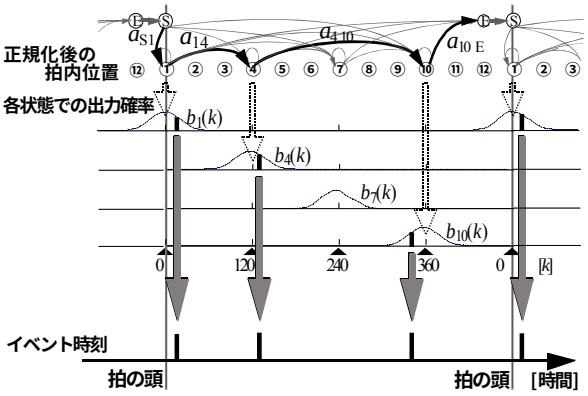
本研究では，1 拍の中の発音時刻を HMM を用いてモデル化し，HMM の隠れ状態を，正規化後の拍内位置(i)に対応づけ，モデルの出力を観測されたイベント時刻(k)，すなわち実際の発音時刻に対応づける（**図 3**）．

遷移は仮想的な状態である Start から始まって，End で終了する．和音など，同じ発音時刻で複数の音が出力される演奏の場合は，同じ状態を自己ループする遷移となる．以下に，遷移の例をあげる．

・ 拍内に 8 分 3 連音符が 3 つ並んでいる場合
　Start → 1 → 5 → 9 → End
・拍内に 16 分音符が 4 つ並んでいる場合
　Start → 1 → 4 → 7 → 10 → End
・拍の頭で 2 音同時に鳴る和音の場合
　Start → 1 → 1 → End
・拍内に音がない場合
　Start → End

HMM の各パラメータを以下のように意味づけた．

・隠れ状態 i： 発音時刻の正規化後の拍内位置（i は 1 から 12 の整数）
・出力 k： 拍内のイベント時刻（k は 0 から 479 の整数）
・状態遷移確率 a_{ij}： 拍内の i の位置で発音した後 j の位置で発音する確率．遷移は，つねに拍内の前から後ろに向かって進むため，a_{iS}=0，a_{Ei}=0 となり， $i>j$ のときは a_{ij} = 0 となる．
・出力確率 $b_i(k)$： 発音時刻の正規化後の拍内位置が i のときに，拍内のイベント時刻が k となる確率．状態 Start と End は出力を出さないため，対応する出力分布は存在しない．状態遷移はつねに Sから始まるため，HMM 初期状態分布 π_i は， π_S=1，π_i (i= 1,2,...,12, E)=0 となる．



12 個の状態のうち，1, 4, 7, 10 以外の状態に関する遷移と出力確率は省略している．

図 3：１拍の隠れマルコフモデルの概略図

3.2 モデルパラメータの教師なし学習

演奏された発音時刻系列 $\boldsymbol{y}$のみから，その状態遷移系列の確率 $P(\boldsymbol{y})=P(\boldsymbol{y}|\boldsymbol{\lambda})$を最大化するようなモデルパラメータ $\boldsymbol{\lambda}$を推定する．HMM では，隠れ状態遷移が観測できないため，パラメータを直接，最尤推定することはできない．そこで，Baum-Welch アルゴリズムを用いて $\boldsymbol{\lambda}$および状態遷移系列の再推定を繰り返すことにより，尤度 $P(\boldsymbol{y}|\boldsymbol{\lambda})$を最大にするモデルパラメータ $\boldsymbol{\lambda}$ を求める．その際，学習データの不足を補うため，ヘルドアウト補間法[13]を用いて $b_j(k)$ を補間する．

3.2.1 Baum-Welch アルゴリズム

Baum-Welch アルゴリズムを用いて，以下のように a_{ij} ，$b_j(k)$ を推定する．推定に用いたのは，3 人の演奏者 A, B，C が，テンポ一定の伴奏に合わせて MIDI ギターでジャムセッションした記録である．各演奏者 2 回ずつ，合計 6 セット（A1,A2,B1,B2,C1,C2）用意した．

・初期値の設定

a^0_{ij} ，$b^0_j(k)$に適切な初期値を与える．a^0_{ij} は A1 から C2 までの 6 つの演奏から求めた a_{ij} の平均値とする．$b_j(k)$は，隠れ状態 i が対応するイベント時刻を中心とする分散σ=20 の正規分布とする．$b_j(k)$ が分散σ=20 の正規分布に近い分布となることは，文献[10]で既に確認している．

・前向き，後ろ向き計算

モデル $\boldsymbol{\lambda}$が与えられたとき， $y_1, y_2,..., y_t$ を出力し，t 番目のイベントで状態 i にいる確率$\alpha_t(i)$ (前向き確率)と，モデルλと t 番目のイベントにおける状態 i

が与えられたとき，t+1 番目以降に $y_{t+1}, y_{t+2}, \ldots, y_T$ を出力する確率 $\beta_t(i)$（後ろ向き確率）とを算出する．

・モデルパラメータ λ の更新

次のような漸化式で， a_{ij} ， $b_j(k)$ を更新する．

$$a^{k+1}{}_{ij} = \frac{\text{状態}i\text{から状態}j\text{へ遷移する回数の期待値}}{\text{状態}i\text{から遷移する回数の期待値}} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a^k{}_{ij} b^k{}_j(y_{t+1} \bmod 480) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \quad (3)$$

$$b^{k+1}{}_j(k) = \frac{\text{状態}j\text{に滞在し}k\text{を出力する回数の期待値}}{\text{状態}j\text{に滞在する回数の期待値}} = \frac{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j) \Big/ \sum_{j=1}^{12} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j) \Big/ \sum_{j=1}^{12} \alpha_t(j)\beta_t(j)} \quad (4)$$

$s.t.(y_{t+1} \bmod 480)=k$

・繰り返し

前向き，後ろ向き計算とモデルパラメータ λ の更新の計算をモデルパラメータが収束するまで繰り返す．

3.3.2 ヘルドアウト補間法

Baum-Welch アルゴリズムでは，一般に繰り返し計算をするごとに，尤度は単調に増加し，モデルパラメータは収束に向かうが，学習データのサンプル数が不十分な場合には，$\alpha_t(i)$や $\beta_t(i)$が 0 になる場合が多くなり，a_{ij}，$b_j(k)$の正しい推定ができなくなる．本研究では，音数の少ない状態 j でも $b_j(k)$が正しく推定できるように，ヘルドアウト補間法を用いて，$b_j(k)$を線形補間する．以下の説明では 1 から 12 までの各状態が対応するイベント時刻を時刻 0 として正規化した時刻 l を用いる(式(5))．

$k-40\cdot(j-1) \geq 0$のとき　$l = k-40\cdot(j-1)$　（5）

$k-40\cdot(j-1) < 0$のとき　$l = k-40\cdot(j-1)+480$

$b_j(l)$を以下のように線形補間する(式(6))．補間係数ζは，$0 \leq \zeta \leq 1$である．

$$\hat{b}_j(l) = \zeta\ \bar{b}(l) + (1-\zeta) b_j(l) \quad (k=0\sim479) \quad (6)$$

ここで，$\bar{b}(l)$は，1 から 12 の状態に対応する各$b_j(l)$の平均の分布である(式(7))．

$$\bar{b}(l) = \sum_{j=1}^{12} b_j(l) \quad (7)$$

式(6)において$\hat{\boldsymbol{b}}_j(\boldsymbol{l})$は，$\zeta$の値を適切に設定すると，最尤推定によって得られた$b_j(l)$よりもよい推定値となることが，スタインのパラドックス[14]として知られている．補間係数ζを求める手法として，本研究ではヘルドアウト補間法(推定法)を用いる．ヘルドアウト補間法は，言語モデルの 1 つである N グラムモデルのパラメータのスムージングにしばしば用いられる手法で，学習データ $\boldsymbol{y}$ を 2 つに分け，片方で最尤推定を行い，もう片方（ヘルドアウト・データ）で補間係数ζの推定を行う手法である．補間係数は EM アルゴリズムに基づいた繰り返し最適化により推定される． ζの再推定式は，次式の通りである．Tはζの再推定に用いた学習データの数である．

$$\hat{\zeta} = \frac{1}{T} \sum_{t=1}^{T} \frac{\zeta \times b_j(l)}{(1-\zeta)\times \bar{b}_j(l) + \zeta \times b_j(l)} \quad (8)$$

本研究では，学習データ $\boldsymbol{y}$ の前半分を Baum-Welch アルゴリズムによる最尤推定に使い，後ろ半分を補間係数ζの推定に使う．ヘルドアウト補間法は，Baum-Welch アルゴリズムの再帰計算で $b_j(k)$が推定されるごとに実行する．

図 4 は演奏 A1 から教師つき学習と教師なし学習で推定した $b_4(k)$の分布を比較したものである．両者は似通った分布となり，教師なし学習でも正しく出力確率 $b_j(k)$が推定できていることが確認できる．

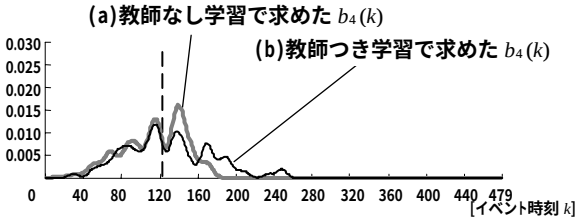


図 4：教師つき/なし学習で求めた $b_4(k)$

3.4 クォンタイズ結果

教師なし学習したモデルパラメータを使ってクォンタイズすることにより，モデルが正しい挙動をしていることを確認する．本研究では，一致率を以下のように定義し，クォンタイズの性能を評価する．一致率は各音の発音時刻が正解と一致している割合である．

$$(\text{一致率}) = \frac{(\text{正規化後の拍内位置が正解と一致した音数})}{\text{正解データ中の音数}} \quad (9)$$

本研究では，閾値処理による機械的なクォンタイズ（**機械**と呼ぶ）と教師データが与えられた条件でモデルパラメータを学習しクォンタイズした結果（**教師つき**と呼ぶ），教師データが与えられない条件でモデルパラメータを学習しクォンタイズした結果（**教師なし**と呼ぶ）を比較した．機械クォンタイズでは，表 1 に示す 3 種類の分解能で一致率を求め評価した．なお，これ以外の分解能では一致率はさらに悪かった．

教師なしの結果は，**教師つき**の結果よりは若干低い一致率となったが，すべての演奏で 7 割以上となり，**教師なし**が高い性能を示すことが確認できた（**表 1**）．

表 1：機械クォンタイズと教師つき／なしの比較

	演奏者 A		演奏者 B		演奏者 C	
	A1	A2	B1	B2	C1	C2
機械（8 分 3 連）	85.6%	67.6%	79.4%	88.6%	57.0%	97.7%
機械（16 分）	37.3%	54.5%	36.8%	34.7%	70.7%	45.5%
機械（16 分 3 連）	48.4%	57.7%	57.8%	51.3%	56.1%	82.5%
教師つき	84.8%	75.9%	80.0%	90.5%	85.1%	95.0%
教師なし	74.7%	78.4%	72.1%	78.2%	84.7%	89.4%

実験で用いた曲の多くは，8 分 3 連中心の演奏であったため，今回の実験では 8 分 3 連の**機械**が高い性能を示すこともあったが，8 分 3 連と 16 分の拍が同程度入っているような演奏や，8 分 3 連と 16 分のどちらの拍が多く入っているか分からない場合には，**教師なし**が有効であった．

4.ボロノイ線図を用いたフレーズ分割

本節では，ボロノイ線図を用いたグルーピング手法[11]を説明し，その手法を応用してフレーズを分割することを提案し，フレーズ分割の性能を評価する．

4.1 ピアノロール上の音符に対するボロノイ線図

図 5 はピアノロール[1]上の音符に対してボロノイ線図を描いたものである．ボロノイ線図とは，ある空間内に幾つかの図形成分が与えられた時に，最近傍則により，その空間を排反な部分空間に分割したときに生じる分割境界の線図形である［15］．ボロノイ線図によって分割された音符を，以下の 2 つのルールに従って，関係の近いものから順に結合していくことにより，階層的なグルーピングができる．

1) ボロノイ線図により囲まれた面積が小さなグループから順に結合する．

2) 結合するグループは，そのグループに最も近い位置にあるグループに結合する．

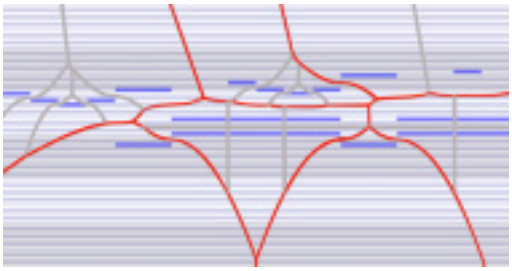


図 5：ピアノロール上の音符に対するボロノイ線図

4.2 フレーズ分割

4.1 のようにしてできた，グルーピング結果のうち，1 つの階層を選択することによりフレーズ分割を行う．ボロノイ線図を用いたグルーピングも，GTTM によるグルーピングと同様に階層構造を持っている．すなわち，グルーピングの一番上の階層では，すべての音がひとつのグループに属し，階層が下になるにつれてグループの数が増加する．したがって，フレーズ分割の際に，どの階層を選択するかによって，各フレーズの長さが変化する．

比較的下にある階層を選択した場合には，1 つ 1 つのフレーズの長さが短くなる．そしてその結果，フレーズを再利用して新たな演奏を生成するとき，フレーズの組み合わせの数が大きくなり，バリエーションの豊富な演奏となる．細かく分割されたフレーズから質の高い演奏が生成可能であることは文献[16]に示されている．しかし，フレーズを細かく分割しすぎると，フレーズに含まれている演奏者固有の特徴が無くなり，個性のないフレーズになる恐れがある．フレーズ分割をする際には，適切な階層を選択する必要がある．

ジャムセッションの場合，多くのフレーズは長さ 1 小節のフレーズを組み合わせて演奏しているため，多くのフレーズが 1 小節前後の長さになるような階層が適切であると考えられる．そこで本研究では，各グループの長さが 1 小節程度になるよう，すべてのグループの平均の長さが 1 小節にもっとも近くなる階層を選択し，フレーズ分割を行うことにした．

4.3 分割結果

フレーズ分割の性能を，以下の一致率で評価した．

$$(\text{一致率}) = \frac{(\text{フレーズが正解データと一致した数})}{(\text{正解データのフレーズの数})} \quad (10)$$

評価に用いたのは，3.4 節でクォンタイズ済みの 3 人の演奏から抜き出した 100 小節である．正解データは手作業で作成した．評価するのは，(1)ボロノイ線図を用いてフレーズ分割した結果，（2）演奏記録を小節線で機械的に分割した結果，である（**表 2**）．

表 2：本手法と小節線での分割結果の比較

	演奏者 A	演奏者 B	演奏者 C
(1)ボロノイ線図による分割	71.5%	70.6%	74.1%
(2)小節線で分割	68.4%	55.2%	78.5%

演奏者 A,B では，（1）が，（2）より高い性能を示した．一方，演奏者 C では，（2）のほうが高い性能を示した．演奏者 C で，（1）が不正解となった部分の演奏の多くは，伴奏のように同じフレーズを繰り返し演奏する箇所であった．このような部分では，小節線による分割のほうが有効となる場合もあったが，多くの場合ではボロノイ線図によるフレーズ分割が有効であることが確認できた．

5．フレーズの再利用

本節では，切り出したフレーズを意図空間上に配置し，再利用する手法として，MDA-OR によるフレーズの空間配置法，局所自己相関関数によるフレーズからの特徴抽出法，正準相関分析による重要な物理特徴量の選択法について述べる．

5.1　意図空間の構成

意図空間は，一対比較によって求めたフレーズ間の主観的類似度に基づきフレーズを配置した空間である．意図空間と出力演奏の物理的な特徴との相関は正準相関分析により求めるが（5.3 節），このときフレーズのサンプル数が多いほうが，よい写像を求めることができる．また，意図空間の次元は，できる限り低いほうが好ましい[9].

---

[1] ピアノロールとは縦軸を音の高さ，横軸を時間とし，音の出るタイミングと鳴り続けている長さを表示するものである．

しかし，この 2 つを同時に満たすことは困難である．なぜなら，フレーズを空間に配置する際にはフレーズ数が増加するに従って，低次元で配置することが一般に困難だからである．たとえば，類似度データを空間に配置する一般的な手法として，Kruskal の多次元尺度法がある．Kruskal の多次元尺度法は，2 つの演奏パターン j と k の類似度を δ_{jk}，多次元空間での距離を d_{jk} とすると，類似度の高い演奏フレーズをど距離が近くなるように多次元空間内の点の位置を決定するものである．（式 11）

$$\delta_{jk} > \delta_{lm} \quad \text{ならば} \quad d_{jk} \le d_{lm} \tag{11}$$

このとき，式(11)が成立する度合いはストレス値 S で表され，S の値が小さいほど類似度をよく反映した空間となる．S の値はデータ数が増加するにしたがって増加し，また，次元数が少なくなるほど増加するため，このような手法で，我々が求めるような空間を構成することは困難である．これは Kruskal の手法が類似度の順序関係を保存しようとするが，データの数が増えるにしたがってそこに無理が生じているためである．

この問題を解決するため，本研究では，フレーズの空間への配置を多次元尺度法の 1 つである MDA-OR(Minimum Dimension Analysis of Ordered Class Belonging)[18]を用いて行う．MDA-OR は，式(11)のような順序関係を完全に成立させることを考えるのではなく，その関係の成立する比率を大きくすることを考えたものである．被験者実験で求めた類似度のように，数量そのものが厳密な意味をもつものではない場合には，漠としたものを漠と扱い，要素の大局的な空間配置をする MDA-OR の手法が適切だと考えられる．

4.3 節で分割した 100 個のフレーズ間の主観的類似度を一対比較により求め，階層的次元作成法により適合度をみながら MDA-OR の次元を上げていった結果，3 次元で適合度が 99.2%となり充分精度が高くなったと考え，空間の次元を 3 次元に決定した．3 軸のうち，最も重要な次元は 1 軸で，次に重要なのが 2 軸である．**図 6** は，100 個のフレーズの Kruskal の多次元尺度法による空間配置と，MDA-OR による空間配置とを比較したものである．Kruskal の手法では，フレーズがほぼ一様な分散で配置されているのに対し，MDA-OR では，1 軸の方向で大きく 2 つのグループに分かれていた．

(a)Kurskal による空間配置　(b)MDA-OR による空間配置

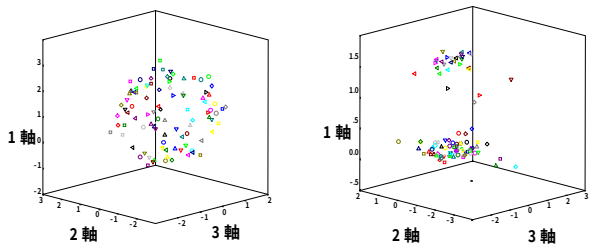


図 6：Kruskal と MDA-OR による空間配置

2 つのグループのフレーズを調べた結果，それぞれソロのようなフレーズのグループと伴奏のようなフレーズのグループであることがわかった．このことは，被験者が類似度判定する上で，フレーズがソロであるか伴奏であるかが重要であるということを表している．本ジャムセッションシステムでは，ソロのフレーズと伴奏のフレーズとの区別はしておらず，仮想演奏者がソロのような演奏を弾くか伴奏のような演奏を弾くかはそのときの意図ベクトルによって決定される．

5.2　局所自己相関関数を用いた特徴の抽出

局所自己相関関数を用いて，フレーズの物理的な特徴を抽出する．ジャムセッションでは，あるフレーズを音高方向，時間方向に平行移動したようなフレーズが使われることがある．特に，伴奏では，コード進行に応じてフレーズ全体の音高を変化させることが多い．このような，あるフレーズとそのフレーズを音高方向や時間方向に少しずらしたフレーズは意図ベクトルが類似していると考えられる．したがって，その 2 つのフレーズからは，同じような特徴量が抽出されることが好ましい．そのようなことを実現する方法として，発音時刻が等間隔となるように時間軸を伸縮したピアノロール上のフレーズ(**図 7**)に対して局所自己相関関数に基づく特徴を抽出した．自己相関関数は平行移動に対して不変であることが知られているが，その高次への拡張が，N 次局所自己相関関数である[19]．ピアノロールの横軸を伸縮したのは，音の長さよりも，音高の変化が特徴として重要だと考えたためである．ピアノロールの縦軸は，ギターの音域を考え，MIDI のノートナンバー36 から 96 までとする．

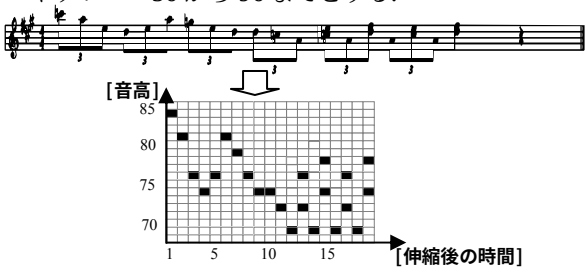


図 7：時間軸を伸縮させたピアノロール

このようなピアノロール上で音が出ているマスを $f(r)$=1，音が出ていないないマスを $f(r)$=0 とすると N 次自己相関関数は，変位方向（a_1,a_2,...,a_N）に対して，

$$x^N(a_1,a_2,...,a_N)=\int f(r)f(r+a_1)\cdots f(r+a_N)dr \tag{12}$$

で定義される．従って，N 次局所自己相関関数は，次数や変位方向のとり方により無数に考えられるが，ここでは次数を 1，局所領域を 5×5 に限定する．局所領域の平行移動により同じになる特徴を除くと，特徴の数は全部で 13 個になる（**図 8**）．

局所自己相関に基づく特徴は，近傍の $f(r)$の積をピアノロール全体に対して足し合わせて得られた値であ

る．本研究では $f(r)$が 1 または 0 なので，その値はピアノロール上で図 8 の各パターンがそれぞれ何回出現したかを数えた値と等価になる．このような特徴は非常に局所的な特徴である．そこで，局所領域の 1 マスに入る範囲を音高方向で 6 種類（1，2，4，6，8，12[半音]），伸縮後の時間方向で 6 種類（1，2，3，4，5，6[伸縮後の時間]）に変化させ，詳細な情報からおおまかな情報までを抽出することを考えた．したがって，合計 468（=13×6×6）個の特徴を抽出した．

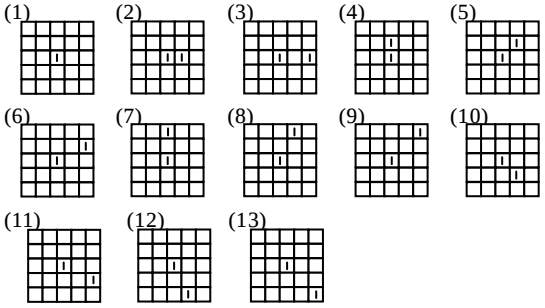


図 8：局所領域の変位パターン

5.3 正準相関分析

5.1 で用いた 100 個のフレーズに対して，意図ベクトルと局所自己相関特徴との正準相関分析を行った．468 個の特徴の中から，一番相関の高い特徴を 1 つずつ選択しながら正準相関分析するという操作を繰り返すことにより重要な特徴量を選択していったところ，50 個の特徴量を選択した時点で，正準相関の値が 0.95 を超え充分高くなったため，そこで終了とした．得られた相関を用いて，新たなフレーズの意図ベクトルが算出可能であることを確認した．

5.4 出力演奏の生成

意図空間上に配置したフレーズから，仮想演奏者のそのときの意図ベクトルに近いフレーズを次々と接続し出力演奏を生成する．その際，フレーズの接続部で違和感がないように考慮して接続する必要がある．

そこで，フレーズを切り出した時にフレーズを囲むボロノイ線図には，もとの演奏記録における前後のフレーズのコンテキストがある程度残っていると考え，フレーズを生成する際に，そのフレーズを囲むボロノイ線図が交差したり，大きく隙間が開いたりしないようにすることを考える．具体的には，前後 2 つのフレーズを囲むボロノイ線図の重なりや隙間がある一定以上にならないように閾値を設定し，閾値を越えたフレーズは選択しないようにした．

各フレーズは，12 小節の 1 コーラス中で，演奏記録からそのフレーズを切りだした位置と同じ位置のみで使用することにする．このようにすることで，コード進行の制約を満たした演奏が生成できる．

仮想演奏者が，演奏を生成する際に意図ベクトルに対応するフレーズが，フレーズデータベース上にない場合には，標準的なフレーズを多数収めてあるフレーズデータベースからフレーズを検索し演奏を生成する．このようにすることで，フレーズを模倣しようとする演奏者のフレーズの数が少ない場合でも，その演奏者のフレーズと標準的なフレーズの両方を用いて演奏を生成することができる．

6．まとめ

本稿では，ジャムセッションの演奏記録から演奏者固有のフレーズを獲得し，フレーズデータベースを生成して再利用するための手法として（1）教師なしクォンタイズ手法，（2）ボロノイ線図を用いたフレーズの分割手法，（3）MDA-OR と正準相関分析を用いたフレーズの再利用法について述べた．（1）では，教師なし HMM クォンタイズの結果が機械的なクォンタイズに比べて，高い性能を示すことが確認した．（2）では，ボロノイ線図を用いたフレーズ分割法が有効であることを確認した．（3）では，獲得したフレーズを意図空間上に配置し，出力演奏を生成するしくみについて述べた．

今後，演奏者による演奏とその演奏者を模倣した仮想演奏者の演奏とを比較し検討していく．

参考文献

[1] 青野裕司, 片寄晴弘, 井口征士：バンドライクな音楽アシスタントシステムについて，情報処理学会研究報告，94-MUS-8，Vol.94, No.103, pp.45-50, 1994.

[2] 青野裕司, 片寄晴弘, 井口征士：アコースティック楽器を用いたセッションシステムの開発，電子情報通信学会論文誌，D- Ⅱ，Vol. J82-D- Ⅱ，No.11, pp.1847-1856,1999.

[3] 和気早苗，加藤博一，才脇直樹，井口征士：テンションパラメータを用いた協調型演奏システムー JASPER ー，情報処理学会論文誌, Vol.35, No.7, pp.1469-1481, 1994.

[4] 近藤欣也，片寄晴弘，井口征士:音楽情報から奏者の意図を理解する伴奏システム JASPER++，情報処理学会第 46 回全国大会，1-373, 7Q-8 , 1993.

[5] 後藤真孝，日高伊佐夫，松本英明，黒田洋介，村岡洋一：仮想ジャズセッションシステム：VirJa Session，情報処理学会論文誌，Vol.40, No4, pp.1910-1921, 1999.

[6] 日高伊佐夫，後藤真孝，村岡洋一：すべてのプレーヤーが対等なジャズセッションシステムⅡ，ベーシストとドラマーの実現，情報処理学会研究報告, 96-MUS-14,Vol.96, No.19, pp.29-36, 1996.

[7] M. Nishijima and Y. Kijima：Learning on Sense of Rhythm with a Neural Network-The NEURO DRUMMER, Proc. of ICMPC, 1989.

[8] M. Nishijima and K. Watanabe：Interactive music composer based on neural networks, Proc. of ICMC, pp.53-56, 1992.

[9] 浜中雅俊，後藤真孝，大津展之: 学習するジャムセッションシステム:演奏者の振る舞いのモデルの獲得，情報処理学会研究報告，Vol.2000, No19, pp.27-34， 2000.

[10] 浜中雅俊，後藤真孝，麻生英樹，大津展之: 発音時刻の楽譜上の位置を確率モデルにより推定するクォンタイズ手法，情報処理学会論文誌, Vol43, No2, pp234-244, 2002.

[11] M. Hamanaka and K. Hirata : Applying Voronoi Diagrams in the Automatic Grouping of Polyphony, FIT2002 情報技術レターズ, LF5, pp.101-102, 2002.

[12] F. Lerdahl, R. Jackendoff: A Generative Theory of Tonal Music, the MIT Press, 1983.

[13] 北研二: 言語と計算ー4 確率的言語モデル，東京大学出版, 1999.

[14] B. Efron and C. Morris: Stein's Paradox in Statistics, Scientific American, Vol.20, pp. 451-468, 1977.

[15] Franz Aurenhammer: Voronoi Diagrams - a Survey of Fundamental Geometric Data Structure, ACM Computing Surveys, Vol. 23, 1991.

[16] D.Cope: Computer Modeling of Musical Intelligence in EMI, Computer Music Journal, Vol.16, No.2, 1992.

[17] J.B. Kruskal and M. Wish: Multidimensional Scaling, Sage Publications, 1978.

[18] 林知己夫: データ解析法の進歩，放送大学教育振興会, 1988.

[19] J.A.Mclaughlin and J.Raviv: Nth-order autocorrelations in pattern recognition, Information and Control, pp.121-142, 1968.