2006－EC－5 (8)
2006／12／16

MIDI シーケンスデータの 2step 打ち込み法への鼻歌による音高入力の適用

伊藤 直樹† 西本 一志†

音楽制作における MIDI シーケンスデータ入力の一手法に，鼻歌入力(Voice-to-MIDI)がある．この入力法の主たるメリットは，覚えたメロディを手動で音名に変換する必要がない点であるが，良好な入力結果を得るのは難しかった．そこで我々は，リズムタッピングを併用することにより，比較的簡易に入力精度を上げる方法を提案する．本稿では，この提案手法と通常の鼻歌入力との比較を行い，提案手法の有効性を評価する．さらに大島，宮川らが提案したメロディを音高とその他の要素に分けて 2 回で入力する 2step 打ち込み法の音高入力に対して本手法を適用するための検討を行った．

A Voice-to-MIDI Pitch Input Method with Concurrently Using Rhythm-Tapping

NAOKI ITOU†, KAZUSHI NISHIMOTO†

Voice-to-MIDI, one of the input methods for MIDI sequence data, has a merit that users can input melodies intuitively, so that they are released from tasks to translate their memorized melodies into chromatic pitches manually. However, the quality of translation by the ordinary VtoM was not satisfactory. Hence, we propose a method to achieve more accurate translations by concurrently using rhythm taps with the VtoM. In this paper, we describe our system and compare our system with the ordinary VtoM. In addition, we discuss to apply our method to the MIDI input method called "Two-Phase MIDI Sequence Input Method," proposed by Ooshima, Miyagawa, et. al.

1.はじめに

本研究では，歌唱による認識精度の高い MIDI データ入力システムの構築を目指しており，本稿ではそのためのリズムタッピング併用歌唱入力法の提案と，この手法を MIDI データの 2step 打ち込み法 [1]の音高入力に適用するための検討を行う．

音楽制作における MIDI シーケンスデータの入力には，鍵盤などの MIDI 規格対応楽器を用いたリアルタイム入力，楽譜・ピアノロールなどを用いたノンリアルタイム入力が存在している．また，スキャナを用いた楽譜認識 [2]なども可能となっている．

しかしこれらの入力法では，特に自作のフレーズの入力やいわゆる "耳コピ"を行いたい場合，つまり楽譜などの音高やリズムが記述された情報がない場合，記憶されたメロディやフレーズから 1 音ずつ自ら音高やリズムを探って変換し，入力する作業が必要となる．この問題を解決する入力手法として，歌で音符を入力する鼻歌入力（Voice-to-MIDI）法がある．

†北陸先端科学技術大学院大学
†Japan Advanced Institute and Science of Technology

鼻歌入力は，音高やリズムの特定をコンピュータが行う．よって，ユーザはマイクに向かって，頭に浮かんだり音楽を聴いて覚えたフレーズを歌うだけで音符を入力可能であり，特に絶対音感や相対音感を持たないユーザにとって，もっとも理想的な入力方法である．

しかし，現在鼻歌入力機能を搭載したソフトはさまざま販売されているが，どのような歌い方をしてもよいわけではなく，よい入力結果を得るためには，歌い方を工夫する必要がある．例えば，「タタタ・・・」のように 1 音毎のリズムの区切りが分かりやすい発音や歯切れよい歌い方，表情を付けず一定に歌うことが推奨されている [3) 4)]．そのためカラオケのようにメロディに歌詞をつけて歌ったり，表情をつけて歌うことには対応できず，実際，誤認識や欠落，余分な音符の入力などが多数発生する．

そこで我々は，歌詞のまま歌ったり，表情をつけて歌うなどのより自然で一般性のある歌唱スタイルでも精度よく認識される歌唱入力システムの構築を行っている．

上述したような既存システムが推奨している歌い方から，鼻歌入力では，1 音ごとの発声開始や発声終了を適切に認識させることが重要で

あると推測できる．そこで本稿では，歯切れよく歌う代わりに，メロディのリズムを区切る情報として，歌唱と同時にメロディのリズム情報をタッピングによって与えることにより，より高い認識精度を実現する手法を提案する．プロトタイプシステムを用いて，リズムタッピング付歌唱入力法の音高取得に関する評価を行う(したがって音長やベロシティなどについては本稿では扱わない)．提案手法と既存の鼻歌入力ソフトとの比較により，提案手法の有用性と問題点を検討する．

加えて，楽器演奏が苦手でも表情のあるMIDI データ作成が可能な手法として提案されながら，これまで鍵盤などを用いて手動による音高変換を行って入力する必要があった 2step MIDI 打ち込み法 [1]の音高入力作業への本手法の適用を検討するため，鍵盤による音高入力との比較を行った．

2. システム構成

2.1 概要

我々が提案する手法はとてもシンプルなものであり，1 音毎の区切りが明確になるように歌う替わりに新たに区切りを示す情報を歌唱と同時に入力することで実現される．具体的には，歌唱するメロディのリズムに合わせて鍵盤楽器や PC キーボード，ボタンなどをタッピングすることにより 1 音毎のリズム区切りを作り出し，これを併せて入力する．

なお本稿で用いるタッピングは，手で拍を打つように打拍後すぐに手を離すようなタッピングではなく，例えばピアノのダンパーペダルのように動作に必要な時間だけ押下し続けるようなタッピングを想定している．

そのため，通常鼻歌入力は，時間の流れに従って繰り返し瞬間的なピッチや倍音構造を算出し，その時系列データの変化を認識することによって音高やリズムを推定するが，提案法では，ボタンなどを押下している間だけ瞬時ピッチ算出処理を行い，ボタンの押下が終了したらその押下中の音高を推定するようになっている．

これは音符の区切りを明示する役割の他に，歌唱時以外に息や咳，喋り声などが誤って変換されることを防ぐ副次的効果もある．

2.2 実装

作成したシステム（図 1）について述べる．

なお本稿では，鍵盤楽器によるタッピングを用い，1 つの鍵のみをタッピング可能な仕様にしたが，複数鍵のタッピングも可能である．

2.2.1 入出力される要素

システムは Microsoft Visual C#2005, 瞬時ピッチ算出部は Visual C++2005 の dll で作成した．また音声録音には DirectSound を用いている．

入力は音声波形と MIDI イベント，出力は E2-G4 までの半音単位の音高（A4 = 440Hz）となり，各種処理はオンライン（リアルタイム）で行う．なお入力音声は 44.1kHz, 16bit, モノラルでサンプリングされ，MIDI イベントは鍵盤楽器（MIDI 対応楽器）から入力される．また，評価実験のために，出力として note on/off や瞬時ピッチ列などの情報の記録，入力波形の録音が行えるように拡張した．

2.2.2 動作概要

システムに MIDI イベントの note on が入力されたらこれをトリガーとして，入力音声波形から瞬時ピッチ算出処理(単位は cent)が note off イベントの入力まで繰り返される．その後瞬時ピッチの時系列から半音単位でその note on～note off 間の音高を推定する．

瞬時ピッチ算出には短時間フーリエ変換 (フレームサイズ= 4096samples, STFT 間隔 = 512samples) とスペクトルの内挿 [5]を用いている．そして note on 入力後，入力波形から最初の FFT フレームを充填しはじめる位置は，充填開始点を S_{start}，波形サンプリング開始～ note on 入力時刻までを ΔT_{note_on} と置くと式 1 で導出される．なお note on 入力時刻の解像度が 1ms なので充填開始点 S_{start} は誤差 44samples の範囲で求まる．

$$S_{start[samples]} = \frac{44100_{[samples]}}{1000_{[ms]}} \times \Delta T_{note\_on[ms]} \quad \text{- 式 1}$$

また瞬時ピッチの時系列から音高を推定する方法は，note on～note off までの区間において半音単位で瞬時ピッチのヒストグラムを求めたときの最頻値を求める方法を用いた．

2.2.3 瞬時ピッチ算出処理の補償

Windows の音声録音の仕組みや FFT フレームサイズの問題により，note on によるトリガーの後，初回の FFT フレームの充填開始～瞬時ピッチ算出結果が得られるまでには，フレーム充填だけでも 4096samples = 約 100ms の大きなディレイが必ず生じる．その結果，速い打鍵を行うと一度も瞬時ピッチを取得できないまま note off が入力され，音高のとりこぼしが発生しかねない．例えば BPM=120 で 16 分音符は 125ms に相当するが，note on ～ note off の時間となると更に短くなり，100ms に満たないことも起こ

りうる．更にこれが連符となると，次の音符の発音タイミングに間に合わせるように打鍵しなければならず，より短い打鍵時間となるため音高認識精度に影響しかねない．

そこで，瞬時ピッチ算出処理に関する 2 つの補償を行っている．1 つ目は note on 入力時刻からオフセットして以前の波形から処理を開始すること，2 つ目は note off 入力時刻からさらに延長して瞬時ピッチ算出処理を行うことである．

1 つ目については，note on 以前の波形が含まれていても note on 後の波形音量が優勢であれば適切に瞬時ピッチを算出できる．様々試した結果現在は，この量を 4096samples 分，つまり FFT フレーム 1 つ分としている．結果として瞬時ピッチ算出などにかかる時間分（コンピュータ性能や瞬間的な負荷量によってこの時間は変わる）のディレイのみとなる．

上記補正の結果，式 1 は，オフセットするサンプル数が 4096samples なので，式 2 のようになる．

$$S_{start[samples]} = \frac{44100_{[samples]}}{1000_{[ms]}} \times \Delta T_{note\_on[ms]} - 4096_{[samples]}$$

- 式 2

2 つ目については，補償する時間は note on～最初の瞬時ピッチ算出結果が得られるまでの時間としている．本来波形が入力された瞬間から瞬時ピッチが算出されるべきであるが，上述したようなディレイが発生する．そこでこのディレイで得られなかった瞬時ピッチ分を補償する．なおこの note off 後の補償については，その時間内に次の note on が入力されれば中止される．

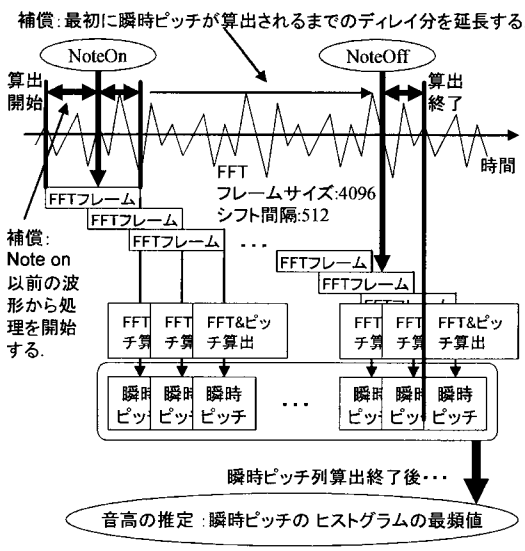

図 1 音高推定アルゴリズムの概念図

3. 評価実験

3.1 実験概要

課題曲を歌詞をつけて歌唱したときの音高認識結果を以下の 3 種のソフトで比較する．

1. 提案手法
2. YAMAHA XGworks ST
3. Internet SingerSongWriter Lite5[6)]

また，2step 打ち込み法の音高入力フェイズへの提案手法の応用について検討するため，鍵盤楽器による課題曲のメロディの入力も同時に行った．

いずれの入力法も誤認識や余分な音高は後から修正するという想定で行ったため，原則的に途中で最初からやり直すことは認めなかった．しかし鍵盤楽器による入力については，やり直しや試し弾きなどを認めた．

被験者は，筆者らが所属する大学の男子学生 6 名である．予備調査により被験者の音楽知識や能力を調べた．内容を以下に示す．

1. 指された鍵の音名を回答：「鍵の音名」
2. 弾かれた単音の音名を回答:「絶対音感」
3. ピアノで弾かれた単音をドとしたときに，相対音階で指定された音を歌い，ハーモニー作成:「相対音感」
4. 「われは海の子」を歌いながら鍵盤をメロディのリズムでタッピングさせる．

項目 1-3 はいずれも全 5 問ある．項目 1-3 について各被験者の 5 問中の正解数を表 1 に示す．

表 1 項目 1-3 の 5 問中の正解数

	鍵の音名	絶対音感	相対音感
被験者A	5(問)	3(問)	3(問)
B	0	0	0
C	5	2	2
D	5	0	0
E	5	1	1
F	5	1	3

また各被験者の音楽歴を以下に示す．

A:エレクトーンを 9 年
B:楽器演奏歴なし
C:ピアノを 5 年,ギターを 10 年
D: 楽器演奏歴なし
E:ドラムとギターを 5 年
F: ボーカル,ギター,ベース,ドラムを 8 年

提案システムの対象は，主に音感を持たないユーザであるが，この実験では，様々な被験者のデータを得るために音感があると思われる被験者にも参加をお願いした．

実験は大学内の防音室を用いて行った．実験環境の構成を図 2 に示す．

記録用に 3 台のコンピュータを用意し，PC1 が XGworksST(以 下 ， XGW) ， PC2 が

SingerSongWriter Lite5.0(以下，SSW), PC3 が提案システム(以下，提案法)を受け持つ．また，鍵盤楽器による課題曲入力（以下，鍵盤入力）については，PC3 上で自作ソフトにより音高情報を記録した．

歌声は Shure SM58 マイクから，鍵盤情報は CASIO LK-80 キーボードの MIDI OUT から取得する．マイク入力された歌唱は Behringer MX802 ミキサーを通じて分配され，PC1,2 でそれぞれオンライン MIDI データ変換を行い，同時に PC3 で鍵盤からのタッピング情報とマージされて提案法によるオンライン音高推定処理を行う．鍵盤入力時は PC3 のみを用いて，鍵盤から入力された MIDI NoteOn 情報の音高を記録する．

なお市販ソフトウェアの個別調整項目については，XGW については，音域を E2-G4(ソフトの表記上は E1-G3)，MIDI 変換時のクオンタイズなし，全音階モードに設定して処理させた．SSW については，特に設定は行わなかった．

選曲は，作曲させたり新規のメロディを覚えさせることは難易度が高いため避け，多くの人が既に知っていると思われる童謡「赤とんぼ」（図 3）を選んだ．童謡の楽譜は発行者によって調が違うことがあるが，今回は野ばら社刊「童謡」に収められた変ホ長調の楽譜に従い被験者に聴取させるメロディのみを MIDI データで作成した．演奏テンポは BPM=80 とした．またこの曲は 31 音符で構成されている．

実験手順は，最初に課題曲を 3 回聴取させ，できるだけ覚えるように指示した．次に歌詞を見ながら覚えたメロディをタッピング付歌唱で 3 回入力させ，次に鍵盤で 1 回入力させた．鍵盤入力では 10 分間の制限を設けたが，10 分以内でも最後まで入力でき満足だと思えば被験者の判断により終了を認めた．そして最後にアンケートへ回答させた．

なおこの一連の過程で楽譜は一切呈示しなかった．また全てのデータ記録は第一筆者が行った．課題曲は被験者 E が一部覚えていないとアンケートで答えたものの,残りの 5 名は全員覚えていた．

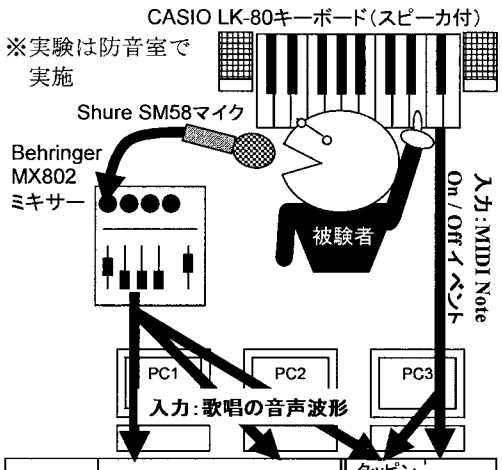

図 2 実験環境および各コンピュータの役割

3.2 実験結果

必ずしも被験者が楽譜通り，あるいはそれを移調した音高通りに歌唱できたとは限らない．ゆえに正しく各システムの性能を評価するためには，誤認識結果が被験者の歌唱の誤りによるものか，システムによるものかを弁別しなければならない．そこで，実験中の歌唱の音響波形を録音し，第一筆者が 1 音毎に音高の特定を行い，その音高と各システムの音高認識結果の比較によって正解個数を割り出し，評価を行った．

各音の区切りはタッピングによって得られた区切りではなく，試聴や波形の目視によっておおよその位置を割り出して用いた．音高は，鍵盤楽器のピアノ音と録音波形を同時に鳴らしながら，ヘッドホンで響きを確認することにより特定した．また適宜ハミングによる響きの確認

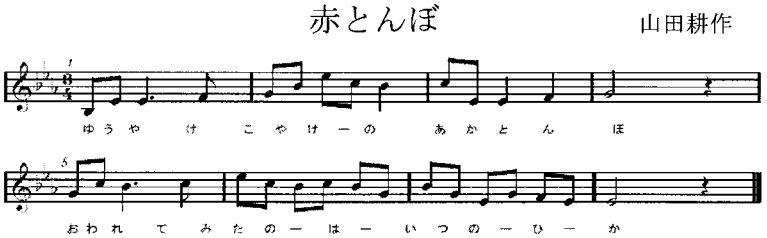

図 3 課題曲「赤とんぼ」

やピッチ抽出ソフトなどを用い総合的に判断した．この作業により各音を，

A. 音高が一意に決まる音
B. 2音の間で決めがたい音
C. 複数音にまたがる音

に分類した．

正解の判定は，割り出した音高とシステム認識結果の音高の比較による．B, C にあてはまる音との比較の場合は，いずれかの音高に該当すれば正解とした．そして，

a. 一致した音
b. 一致しなかった音
c. 1 音を複数音に認識し，かつそのいずれの音も正解と一致しなかった音
d. 本来の 31 音に対して欠落した音
e. 余分な音

に分類して個数を集計した．b. 一致しなかった音については，さらに 2 音の半音単位の音程によって分類した．また，a～d を合計すると 31 音となる．入力 3 回分計 93 音について被験者ごとに集計を行った結果を正解率とともに表 2 に示す．

表 2 より，いずれの被験者とも提案手法の正解率の高さと音符抜けの少なさが伺える☆．正解した音符数については，両側 t 検定では，提案手法が XGW，SSW それぞれに対して 0.1%以下で有意であった．音符抜けの個数については，両側 t 検定では，提案手法が XGW，SSW それぞれに対して 0.1%以下で有意であった．

被験者 E の提案手法の音符抜けの個数は 22 個と多いが，これは例えば「いつのーひーか」は本来 7 音であるが，音を伸ばす箇所を省いて「いつのひか」と 5 音分しかタッピングされていないためである．

余分な音符（音符過多）については，提案手法ではミスタッチしない限り原理的には発生しないため，妥当な結果と言える．なお余分な音符は歌唱中についてのみの集計であるが，既存システムでは歌唱と歌唱の合間などの発声練習や喋り声も変換され，実際にはさらに多くの余分な音符が入力されていた．

一方で提案法は，12 半音差つまりオクターブの誤認識が多く見られた．これは瞬時ピッチ算出アルゴリズムに起因すると考えられる．今回は，オンライン処理の負荷を考え，単純に最も強いピークを基本周波数とみなす方法を用いたが，この方法では，声質や発音などによって倍音成分の方がピークが強かったり，FFT の周波数解像度の問題により，隣あった 2 つのピークに分散されてパワーが弱まってしまうなどの問

表 2　被験者ごとに表した各ソフトの音高認識結果

		音程が0半音差（個）［正解］	1半音（個）	2半音（個）	3半音（個）	4半音（個）	5半音（個）	6半音（個）	7半音（個）	8半音（個）	9半音（個）	10半音（個）	11半音（個）	12半音（個）	13～半音（個）	複数（個）	音符抜け（個）	音符過多（個）	正解率（%）
被験者A	提案	52	19	4	5	3	0	0	0	0	0	2	1	7	0	0	0	0	55.91
	XGW	34	16	5	2	1	0	0	0	0	0	0	0	0	0	1	35	10	36.56
	SSW	8	26	6	3	2	2	1	0	0	0	0	0	0	0	1	44	0	8.60
B	提案	67	1	0	2	3	0	0	2	0	0	2	0	14	0	0	2	0	72.04
	XGW	27	13	2	1	0	0	0	0	0	0	0	0	0	0	0	50	1	29.03
	SSW	18	11	8	4	2	0	0	0	0	0	0	0	0	0	0	50	0	19.35
C	提案	66	7	5	4	0	0	1	0	0	0	0	0	2	0	0	8	0	70.97
	XGW	27	28	3	0	0	0	0	0	0	0	0	0	0	0	0	35	30	29.03
	SSW	2	6	9	8	8	3	4	0	3	1	0	2	0	4	12	31	18	2.15
D	提案	52	13	2	4	1	0	0	0	1	0	1	5	14	0	0	0	0	55.91
	XGW	21	17	16	3	0	0	1	0	0	0	1	0	0	0	3	31	17	22.58
	SSW	8	12	7	4	4	4	0	0	0	0	0	0	0	0	0	54	0	8.60
E	提案	50	11	4	2	2	0	0	0	0	0	0	0	2	0	0	22	0	53.76
	XGW	9	14	11	0	2	0	0	0	0	0	0	0	0	0	1	56	2	9.68
	SSW	14	9	7	9	1	1	0	0	0	0	0	0	0	0	0	52	0	15.05
F	提案	81	4	3	1	0	1	0	0	0	0	0	0	1	0	0	2	0	87.10
	XGW	16	22	2	2	0	0	0	0	0	0	1	0	0	0	0	50	2	17.20
	SSW	7	9	11	7	3	2	0	0	1	0	0	0	0	0	1	52	0	7.53

※音程が 0 半音差～音符抜けまでの合計は 93 音(31 音*3 回分)になる．
正解率＝(音程が 0 半音差の音符個数) / 93 * 100
※複数：1 音が複数の音符に認識され，かつそれらの中に正解の音高がない場合
※音符過多：1 音が複数の音符に認識されたときなどの余分な音符数

☆ XGworks ST や SingerSongWriter Lite5 では，推奨されている歯切れよい歌い方や「タタタ・・・」のような歌い方であれば，認識精度は向上すると思われる．

題に対処できず，倍音成分を基本周波数とみなされる可能性がある．これについては今後ケプストラム分析などの導入を考えているが，逆に考えると提案手法は，ヒストグラムによる音高推定処理と合わせて，非常に簡易な処理法でも高い音高認識精度を実現していると言える．

3.3 2step 打ち込み法への応用の検討

2step 打ち込み法で従来音高入力に用いられてきた鍵盤入力を拡張する形で，提案手法を歌唱による音高入力として応用するための検討をアンケート結果を交えて行う．表 3 に鍵盤入力による各被験者の課題曲「赤とんぼ」の入力終了までの時間と総入力音高数を示す． なおリズムタッピング付き歌唱入力の入力時間は，各被験者とも 1 回あたり 14～18 秒程度であった．

楽器経験者である被験者 A,C,D は時間内に入力を終えている．彼らは提案法について，

A:「鍵盤を押すタイミングが混乱した」
C:「歌のリズムと鍵盤のリズムを合わせるのが難しい」
D:「フィードバックがないから正しく歌えているか不安」

というように評価している．A,C はともにリズムタッピングのやりづらさを挙げている．これについては，被験者 F も

「テンポを保つために発音以外でも押してしまう」

と述べている．これは楽器経験者がテンポの拍打ちにタッピングを用いることに慣れており，メロディのリズムに合わせてタッピングすることに負荷を感じたのではないかと思われる．

一方で，楽器演奏経験のない被験者 B,D は 10 分の制限時間内に最後まで入力できなかった．アンケートの自由記述欄では，被験者 B は，

「リズムを取りながら歌えるので，それほど歌のみと感覚的には変わらなかった．」

また被験者 D は，

「普段意識的に行う行為ではないが，特に違和感はない．感覚的なものではあるがリズムをきざみながら歌ったので若干歌いやすかったようにも思う．」

と回答した．ともにメロディのリズムタッピングを行うことに違和感を感じずに受け入れていることが分かる．この 2 名だけでは判断しかねるが，提案法は 2step 打ち込み法が対象とする楽器演奏未経験者や苦手な者には有用である可能性が高い．

また入力音高数は，ある程度音感を持つと思われる被験者 A 以外は，非常に膨大な量となった．ゆえに後の削除・編集作業の手間を考えると，提案法はメロディのリズムタッピングに慣れることによって，楽器経験者にも有用だと思われる．

表 3 鍵盤による「赤とんぼ」の
メロディの音高入力

	入力終了までの時間	総入力音高数	楽器経験
被験者A	24秒	38音	エレクトーン9年
B	timeout	724音	なし
C	6分55秒	529音	ピアノ5年他
D	timeout	1063音	なし
E	9分54秒	922音	ドラム5年他
F	timeout	954音	ボーカル8年他

4. 結論

提案システムでは，リズム区切りを明確化した特殊な歌唱法を用いずに歌詞をつけて歌唱しても高い音高入力精度を得られることが分かった．2step 打ち込み法の音高入力への提案法の応用を検討し，この打ち込み法が対象としている楽器演奏未経験者にはよい評価を得た．一方で，楽器演奏経験者にはタッピングに負荷があることも分かった．

5. 今後の予定

まずより安定した瞬時ピッチ算出アルゴリズム，音高推定アルゴリズムの検討を行う．また本稿では 2step 打ち込み法の音高入力に対する歌唱入力の適用可能性について検討を行ったが今後，実際に 2step 打ち込みシステムへの組込みを行い，システム全体としての評価を行う予定である．そしてタッピングによる音長入力が可能かを含め，リズムタッピング付歌唱入力システムの構築を目指す．

謝辞

本研究は，科学技術研究費補助金基盤研究(C)(2)課題番号 16500580 の支援を受けて実施したものである．

参考文献

1) 大島千佳，西本一志，宮川洋平，白崎隆史: 音楽表情を担う要素と音高の分割入力による容易な MIDI シーケンスデータ作成システム, 情報処理学会論文誌, Vol44, No.7, pp.1778-1790, 2003.
2) 株式会社河合楽器製作所: スコアメーカー, http://www.kawai.co.jp/cmusic/products/scomwin/.
3) ヤマハ株式会社: XGworks ST, http://www.yamaha.co.jp/product/syndtm/p/cmp/xgwstw/index.html.
4) 株式会社メディア・ナビゲーション: 鼻歌ミュージシャン 2, http://medianavi.co.jp/product/hana2/hana2.html
5) 原 裕一郎，井口 征士: 複素スペクトルを用いた周波数同定: 計測自動制御学会, pp718-723(1983).
6) 株式会社 インターネット: SingerSongWriter Lite5,http://www.ssw.co.jp/products/ssw/win/sswlt50w/index.html