

メロディリズムのタップを併用する Voice-to-MIDI 変換手法の音高 変換精度評価

伊藤 直樹[†] 西本 一志[†]

計算機を用いた音楽制作における MIDI シーケンスデータ入力法のひとつに鼻歌入力法がある。しかし既存システムでは 1 音毎の区切りがうまくゆかないことによる変換精度低下が起こる。この問題に対して我々はタップ併用型 Voice-to-MIDI 手法を提案し、既にタタタ歌唱を前提とする既存 VtoM システムとの比較実験を行い、既存システムと比較して勝るとも劣らない精度で音高変換できることを示している。しかし、歌詞歌唱などの任意の発音の歌唱を許容する既存システムとの比較はこれまで行っておらず、本システムの有用性を十分に示すことができていなかった。そこで今回タタタ歌唱を前提としない、自由歌唱可能なシステムとの比較を実施し、本システムの有用性を明らかにした。

Evaluation of Pitch Translation Accuracy of a Voice-to-MIDI That Concurrently Uses Rhythm Taps of Singing Melody

NAOKI ITOU[†] KAZUSHI NISHIMOTO[†]

Voice-to-MIDI is an input method of MIDI sequence data just by singing a melody. However, the quality of translation of the ordinary Voice-to-MIDI systems is insufficient. One of the most significant problems is the poor accuracy of the segmentation of notes. To solve this problem, we already proposed a novel Voice-to-MIDI method that uses concurrently input rhythm tapping while singing. We confirmed that our prototype system achieved much more accurate translation results than that of the ordinary system that imposes users to sing in a special way called "tatata singing." However, we haven't yet compared our system with a system that allows the users to sing in any ways like singing with lyrics. Hence, in this paper, we conducted experiments to compare our system with the system that allows free singing and we confirmed the superiority of our system.

1. はじめに

計算機を用いた音楽制作における MIDI (Musical Instrument Digital Interface) シーケンスデータ入力法のひとつに、鼻歌入力 1)-3) (Voice-to-MIDI: 以下 VtoM) 法がある。VtoM を使うと、ユーザは、マイクに向かって頭に浮かんだメロディや記憶しているフレーズを歌うだけで音符を入力できるので、特に絶対音感や相対音感を持たないユーザや楽器演奏技術の無いユーザにとって有用な入力方法である。しかしながら、従来の VtoM システムには多くの課題があった。

VtoM システムの処理は、一般に

- 歌唱区間の検知
- 1 音毎の区間検知
- その区間のピッチ採集
- そのピッチ情報からの区間音高推定

という手順で行われる。この各処理ステージで得られた結果は、いずれも連鎖的に次の処理の結果に影響を

与える。したがって初期の処理ステージでの誤りは、それ以降のステージでのさらなる誤りを引き起こし、最終的に得られる変換結果をきわめて精度の悪いものとしてしまう。これを防ぐためには各ステージにおいてできるだけ高い精度の処理結果を出すことが必要となる。とりわけ、初期のステージである歌唱区間の検知および 1 音毎の区間検知の精度を上げることは、それ以降の処理ステージへの波及効果が大きいので、極めて重要である。

ところが、歌唱区間や 1 音毎の区間を計算機処理によって検知することは容易ではない。このため、多くの既存 VtoM システムでは、すべての音を「タ」という音で明確に区切って発声して歌う「タタタ歌唱」のような、特殊な歌唱方法が求められる。これにより一定の水準の処理結果が得られるようになる。しかしながら、たとえば初めに歌詞を作ってからメロディ作曲する「歌詞先作曲」の場合、歌詞の持つイントネーションなどがメロディに大きく影響するため、歌詞をそのまま歌唱することが不可欠である。このような場合、歌唱スタイルを制限せず、任意のスタイルの歌唱

[†] 北陸先端科学技術大学院大学

によって MIDI シーケンスデータを入力することができる VtoM システムの実現が求められる。

そこで、我々はタップ併用型 Voice-to-MIDI (以下 TVM と略す) 手法を既に提案した 4)。これは、計算機が苦手とするが人にとっては容易な区間区切り作業を人が担当し、計算機は得意だが人が苦手とするピッチ抽出を計算機が担当する、人と計算機の協調型システムであると言える。TVM を用い、タタタ歌唱を前提とする既存 VtoM システムとの比較実験を行い、TVM が既存システムと比較して勝るとも劣らない精度で音高変換できることを示した 5)。

しかし、歌詞歌唱などの任意の発音の歌唱を許容する既存システムとの比較はこれまで行っておらず、本システムの有用性を十分に示すことができていなかった。そこで今回タタタ歌唱を前提としない、自由歌唱可能なシステムとの比較を実施し、あらゆる歌唱スタイルにおける TVM の優位性を実証したので報告する。

2. 先行研究

文献 6)7)では音声認識のために、本研究と同様に発声に併せたタッピングなどによる区切り情報入力を行っている。これらにより音節区切り情報の効果は示されているが、V-to-M システムへの適用を目的とした研究ではない。またこれらの文献より、TVM の歌詞認識への応用も考えられるが、歌詞認識の難しさ 8)もあり、現時点では研究の対象とはしていない。

VtoM の精度向上に関する文献 9)では、音程の外れた歌唱にも対応可能な手法についても述べられており、発声した個々の音が絶対音高から外れていても、相対音高としてはスケールを構成していることを利用して、補正を行うことが提案されている。また文献 1)の鼻歌入力システムでは、スケール上の音に優先して認識されるように重み付けを行うことが可能である。これらの音高認識結果の補正手法は、音響処理レベルを超えた、より高次の音楽処理レベルの処理ステージで適用される技術であり、TVM と組み合わせることによりさらに高精度な V-to-M システムを実現することが可能と考えられる。

3. タップ併用型 Voice-to-MIDI システム

3.1 既存 VtoM システムの問題点

既存の V-to-M システムに歌詞歌唱を入力したときの問題点を示す。市販の V-to-M システムに童謡「赤とんぼ」(野ばら社刊「童謡」の変ホ長調版 10)を使用: 図 1) を歌詞歌唱入力した結果を 2 例示す。

図 2 にタタタ歌唱入力を前提とするある市販システムにおける「(ゆうやけこや) けえのあかとんぼ」部分の変換結果を示す。上段は入力された歌詞歌唱の音声波形を、中段は正解のメロディラインを手動入力して 2 オクターブ移調したもの(正解データ)、下段は V-to-M システムによる認識結果を示す。このシステムは音量変化によって音が区切られると推測されるが、本来 1 音であるのに複数の音に認識されてしまったり、逆に複数の音に分割されなければならない箇所が 1 音と認識されてしまったりしている箇所が多数ある。

図 3 は、別のシステムによる「おわれてみた」部分の変換結果である。このシステムでは主に音高変化によって音が区切られると推測されるが、意図しないピッチの変化にも反応してしまい、「お」と「て」の部分で余計な音が入力されてしまっている。



図1 赤とんぼの楽譜

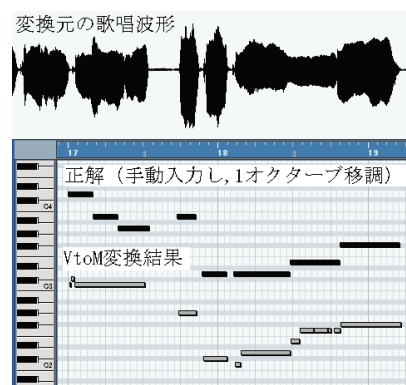


図2 音量によって区切られ、複数音が 1 音に、1 音が複数音に変換された例(赤とんぼの「けえのあかとんぼ」)



図3 音高変化によって区切られ、余計な音が入力された例(赤とんぼの「おわれてみた」)

このように、従来の V-to-M システムは歌唱音声デ

ータを適切に 1 音ずつに区切れず、その結果個々の音の音高や音長の誤認識が起こっていると言える。

3.2 タップ併用型 VtoM 手法の概要

上記のような問題に対処するためには、音量の変化が乏しいことによって音が区切られない問題やピッチの変化による意図しない区切れの発生を同時に抑えられなければならない。そこで TVM では、機械が苦手な音符区切り判定を人間が手動処理し、人間の苦手なピッチ抽出をシステムが自動処理するという協調的な処理手法を採用した。

ユーザは、歌唱と並行してメロディの各音を区切る情報（リズム区切り情報）を入力する。具体的には、歌唱するメロディのリズムに併せて鍵盤楽器や PC キーボード、あるいはなんらかの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入力する。その上で、鍵やボタンが押下された時点から短時間ピッチ算出処理を開始し、鍵やボタンが離され押下が終了した時点か歌唱の途切れが検知された時点のどちらか長い方まで短時間ピッチ列算出を継続し、この間を 1 つの音符に対応する音声データであるとして、得られた短時間ピッチ列から 1 つの音高を推定し出力する。

3.3 プロトタイプ構成

上記の処理を実装した TVM プロトタイプシステムについて述べる。入力は音声波形とリズム区切り情報、出力は D2-F5 までの半音単位の音高（A4 = 440Hz を基準とする）である。入力音声は 22050Hz, 16bit, モノラルでサンプリングされる。リズム区切り情報には MIDI キーボードや PC キーボードの打鍵および離鍵の入力時刻情報を用いる。PC キーボードの場合は、タッピングに '<' と '>' の 2 キーを使用し、1 キーのみ連打しても 2 キーを交互に打鍵してもよい仕様とした。処理はオンライン（リアルタイム）で行われる。

キーを押下することにより、システムに打鍵情報（MIDI note on message）が入力されたら、これをトリガーとしてマイクより入力されてくる歌唱音声データからの短時間ピッチ算出処理を開始し、キーが離され離鍵情報（MIDI note off message）が入力されるか、後述する無発声検知機構によって終了が検知されるまで短時間ピッチ算出処理を繰り返し、短時間ピッチの時系列データを記録する。短時間ピッチ算出は、入力波形に対する短時間フーリエ変換(STFT, フレームサイズ $twin = 2048samples$: 約 100ms, フレーム移動間隔 $\Delta t = 128samples$: 約 6ms)から求めたパワースペクトルの D2-F5 相当の周波数間に存在するピークのうち、このパワースペクトルに対する IFFT から求めた循環

自己相関の正の最大値近傍の周波数のものを用いて求める。更にスペクトルの内挿 11)を用いて cent 単位で音高推定を行い短時間ピッチとして出力する。これは周波数解像度不足を補うためである。

離鍵後、短時間ピッチ時系列データから半音単位でとったヒストグラムを生成し、最も頻度の高い音高の音名を求め、これをこの区間に対応する音符 1 つ分の音高として出力する。

3.4 無発声検知機構

以前のシステム 5)6)では、タップを終了することで音長が決定されるシンプルな仕組みであったため、1 音の長さ分だけキーを押下し続けずに、タップしてもすぐに離してしまうようなタップでは十分な量の短時間ピッチ情報が取得できずない問題があった。この点を踏まえて、本システムでは歌唱区間の途切れを検知する機構を実装した。具体的には、本システムではピッチ抽出に循環自己相関を用いているため、タップ後に D2-F5 の音高範囲内に最大の正相関値がなくなれば歌唱区間の終了と判断する。

タップ終了と歌唱終了のタイミングによって終了位置は以下の 3 パターンに分かれる。

- タップ終了後に歌唱が終了：歌唱終了時点
- 歌唱が終了しないまま次のタップ開始：次のタップ開始直前（レガート音）
- タップ終了より先に歌唱が終了：タップ終了時点

この手法により対象とする音高範囲内に目立つ音がなければ、音量閾値などの手法を用いずに有音 / 無音を判別可能となり、周期性がはっきりとした音が存在していなければ環境音の音量変化への動的対応や小音量下でも判別が可能となるなどのメリットがある。

一方でこの手法では、タップ終了後でも、歌唱以外の音に反応したことによって範囲内に最大の正相関値が出現していれば消音されない可能性があるが、PC 内蔵マイクやヘッドセットマイクなど数種類のマイクで調査したところ、概ね良好に作動した。なお、タップ開始～200ms までは無発声を検知しないようにした。また、音が鳴っているにもかかわらず音高範囲内にピッチが無いと判定されることを想定し、音量（パワースペクトルの合計値）が直前のフレームの音量の 90%以上であれば終了しない仕様とした。

4. 評価実験

4.1 実験概要

リズム区切り情報追加による効果と問題点を探るため、歌唱音声データの分割区間数の精度と、各区間の

音高認識精度の評価を行った。すでに「タタタ歌唱」を推奨するシステムを用いていた比較評価を行い、TVMがこのシステムと比較して勝るとも劣らない精度で音高変換できることを既に示している6)。今回は、TVMと同様に歌詞歌唱などの自由な発音の入力を許容するVtoMシステムと比較する。これは、我々の目指す歌唱スタイルを制限しない入力という目的により近い既存システムと考えられる。

なお歌唱の音の立ち上がりおよび立ち下がりも正確に判定するのは困難であるため、今回の実験では、音長やリズムの精度については評価しない。

4.2 楽曲

歌唱する楽曲は以下の2種類である。

- 課題曲（赤とんぼ）
- 各被験者が選んだ自由曲（歌詞のあるメロディを1コーラス程度）

赤とんぼは、音高の範囲が広く、変化も激しいが一方で同一音高が連続する箇所もあり、適度な難しさを持っている。かつ多くの人が知っている曲であることから課題曲に採用した。歌唱テンポによって大きく2種類の歌唱条件を設定し、「テンポ自由」では、被験者の好みのテンポで歌唱させた。また、赤とんぼは通常遅いテンポで歌唱されるため、「BPM=120」で歌唱させ、歌唱とタップの同期が速いテンポでも可能かを検証した。

自由曲では、赤とんぼよりもリズムや音高変化が複雑でより実践的な曲への対応が可能かを検証するために、各被験者自身が選曲したポップスなどのメロディを歌唱させた。

4.3 機材設定

比較に用いた既存 V-to-M システムは、KAWAI: Band Producer 2 12)に付属の鼻歌入力機能(以下、BP2と略す)である。この機能は、予め設定した音量閾値を超過したときと半音単位の音高閾値を超えたときに音符が区切られると変換結果から推測されるが、例えば音量で区切られなかったとしても音高変化があれば区切られるため、歌唱の発音により影響されにくいと思われたため比較対象として採用した。

次にデータの記録および処理手順について述べる。

被験者に試唱させて BP2 の録音音量閾値を設定した後、BP2 に歌唱をリアルタイムで入力し、MIDI データに変換する。同時にその歌唱は Wave 波形として BP2 上で録音される。TVM のためのタップデータの記録については、被験者に歌唱と同時にタップを入力させ、BP2 とは別の PC で記録する。このタップデー

タに BP2 で記録した波形と組み合わせてオフライン処理で MIDI データに変換する。実験では両システムで完全に同じ歌唱波形を使用するために便宜上、本来オンライン処理である TVM をオフライン処理とした。しかし、この実験のために更なる精度向上を目的としたような処理は追加せず、同等の出力結果となる。

なお、BP2 で記録した歌唱波形と TVM のタップデータの同期が必要となるが、TVM 用の PC で歌唱波形をタップと同期させて記録しており、その波形と BP2 の波形を目視して同期位置を探した。

タップに用いたデバイスは、HP: 2710p ノート PC のキー “<” および “>” である。これらのキーは隣接して存在する。被験者は、これらのキーの両方あるいは片方のみを好みに応じて用いる。

4.4 被験者

被験者は、筆者らが所属する大学の男子学生 8 名と女子学生 1 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。項目を以下に示す。

- [1] 「鍵の音名」: ピアノ上の鍵の音名回答
- [2] 「音高聴取」: ピアノで弾かれた単音の音名回答
- [3] 「音の高低」: ピアノで弾かれた 2 音の高低回答

項目 1-3 はいずれも全 6 問ある。各被験者の 6 問中の正解数と楽器経験を表 1 に示す。

なお TVM の支援対象は、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和音楽器経験者・リズム楽器経験者や音感があると思われる被験者にも参加をお願いした。その結果、楽器経験なし 4 名と経験あり 5 名となった。

4.5 実験手順

実験は大学内の防音室を用いて 1 名ずつ行った。

まず VtoM の練習および歌唱しながらタッピングする練習を 5 分ずつ行った後、以下の順序で実施した。まず、被験者に課題曲の童謡「赤とんぼ」の 1 番(全 31 音符: 図 1 参照)を、歌詞を見ながら 3 回聴取させ、メロディをできるだけ覚えるように指示し、

- [1] 赤とんぼ: テンポ自由
- [2] 赤とんぼ: BPM=120
- [3] 自由曲

の順に歌唱させた。この 3 歌唱課題それぞれにおいて表 2 の歌唱条件をランダムな順番で呈示して歌唱させた。赤とんぼについては、それぞれの入力方法について、3 回ずつ歌唱を入力させた。自由曲については、被験者の負担を考えて 1 コーラス程度を 1 回歌唱させた。各被験者の自由曲を表 3 に示す。実験は全て歌詞歌唱(途中で歌詞が分からなくなった場合は適当な発

音でもよい)で行い、実験中は、歌詞カードは見てもよいが楽譜は一切呈示しなかった。

表1 各被験者の予備調査項目 1-3 の正解数と楽器経験

被験者	音名	音高聴取		音の高低	楽器経験
		正解	半音差		
A	6	0	1	5	なし
B	3	0	0	2	なし
C	6	1	0	5	なし
D	3	1	0	6	なし
E	0	1	0	6	太鼓, ムックリ 1 カ月
F	5	0	0	5	和太鼓 2-3 年
G	6	0	0	6	電子オルガン 2 年
H	6	0	4	6	電子オルガン 3 年 ピアノ 5 年
I	6	5	1	6	ピアノ 10 年以上

表2 実験で用いた歌唱条件の組合せ

[A] 赤とんぼ

テンポ	タップ
自由	あり
	なし (BP2 のみ使用)
BPM = 120	あり
	なし (BP2 のみ使用)

[B] 自由曲

テンポ	タップ
自由	あり

注1. テンポ

- ・自由: 好みのテンポで歌唱。
- ・BPM=120: BPM=120 のメトロノームに合わせて歌唱。

注2. タップ

- ・あり: タップしながら歌唱。
- ・なし: 歌唱のみ。BP2 におけるタップの有無による比較用。

表3 各被験者の自由曲

被験者	歌手名	曲名
A	Mr. Children	Over
B	井上あずみ	さんぼ
C	フォーククルセダース	11月3日
D	スピッツ	チェリー
E	Acid Black Cherry	愛してない
F	ブルームオブユース	ラストツアー
G	チャーリー・コーセイ	ルパン三世 その1
H	SMAP	世界で一つだけの花
I	高橋洋子	残酷な天使のテーゼ

4.6 評価方法

被験者が必ずしも楽譜通り、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音高認識性能を評価するためには、楽譜に記載された音との食い違いが被験者の歌唱の誤りによるものか、システムの誤認識によるものかを弁別しなければならない。そこで、BP2 で記録した実験中の歌唱音響波形から、第一筆者が1音毎に音高の特定を行い、これを「正解データ」とした。つまり、楽譜上に記載されている音高ではなく、実際に歌唱された音高を正解データとする。これにより、被験者の歌唱誤りをシステムの誤りとみなしてしまうことを回避し、純粋にシステムの性能を評価できる。こうして得られた正解データ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て評価を行った。

歌唱からの音高特定の手順（正解データの求め方）は以下の通りである。波形処理ソフト（Adobe: Audition1.0）上で、各音の発音開始～終了までをループ再生した音に対して、ピッチを細かく調整可能なピッチベンドホイール付きのキーボード（Ensoniq: MR-76）で音高特定を試みる。もし、ここで決められない場合は、その発音区間内で発音長に応じて適当に選んだ1～4箇所程度のそれぞれについて、ある程度定常な音になるように30～300ms程度の短い範囲でループ再生して局所的に音高特定を行う。あまりにも音高の変化が大きい音や音高の特定が困難な音は評価から除外した。なお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、試聴や波形の目測によっておおよその位置を割り出した。この作業により各音を、

- 音高が一意に決まる音
- 2音の間で決めがたい音
- 発音中に音高が変化する音

の3種類に分類した。なお、BとCに分類される音は、可能性のある音すべてを正解データとみなした。

次に発音開始および終了位置に基づき、個々の音について正解データと認識結果を対応づけ、両者の音高を比較することにより正解を判定した。ここで分類B、Cにあてはまる音との比較の場合は、複数の正解データのうちのいずれかの音高と一致すれば正解とし、

- 正解音：一致した音
- 誤り音：一致しなかった音
- 欠落音：欠落した音

※自由曲では欠落した音を以下に分けて示す。

- 欠落した音の全体数
- 欠落した音の内、他の音と結合された音

[4] 余分音：余分な音

に分類して個数を集計した。自由曲の「3. 欠落した音」については、出力されなかった音の全体数およびその内の正しく区切られず前の音と結合されてしまった音の数についても示す。「4. 余分音」に分類されるのは、本来1つの音が複数音に認識され、かつその中に正解と一致した音があった場合に正解音に加算される1音分を除いた残りの音、および歌唱中における咳等のノイズによるものなどとなる。1~3の音数の合計は、各メロディの全音符数と一致する（赤とんぼの場合31音）。

最後に上記の分類結果を用いて変換精度を求める。例えば、正しく音高が変換された音数が多いが余分な音も多く出力された場合、よいシステムとは言い難い。そこで、歌唱された音数に対して正しく音高が変換された音数の割合を測る再現率、およびシステムが認識した全音符数に対して正しく音高が変換された音数の割合を測る適合率の2つの尺度で評価する。また再現率と適合率を総合して評価する指標としてF値も求める。それぞれ以下の計算で求められる。

- 再現率(%) = 正解音数 / 全歌唱音数 * 100
- 適合率(%) = 正解音数 / (正解音数 + 誤り音数 + 余分音数) * 100
- F値 = (2 * 再現率 * 適合率) / (再現率 + 適合率)

なお全歌唱音数は以下のように求める。

$$\text{全歌唱音数(音)} = \text{正解音数} + \text{誤り音数} + \text{欠落音数}$$

5. 評価実験結果および考察

評価実験結果および考察について述べる。なお、BP2で全体的に欠落音が多い点については、同一音高の連続箇所など複数音が1音に変換されたことが影響することはあるが、その分を除いてもなお大量の欠落音が残る場合がある。そこで音量閾値設定の影響が考えられたため、閾値を調整して検証してみたが変換結果に大きな変化は見られなかった。また音量が小さい音出力された一方で、その音よりも音量が大きい音が欠落したケースも見られたため、原因の特定は困難として断念した。

5.1 赤とんぼ:テンポ自由

「テンポ自由、歌詞歌唱、タップあり」の歌唱条件による入力3回分計93音について被験者ごとに集計を行った結果、およびBP2におけるタップの有無による精度比較用に「テンポ自由、歌詞歌唱、タップなし」の結果を表4-Aに示す。

TVMは、被験者Cの誤り音が多少多いものの、全

体的に欠落・余分音は非常に少なく上手くタップによる音区切りおよび音高変換がなされていると言える。

一方BP2は誤り音が少なく認識した音の音高変換精度は非常に高いものの、欠落・余分音が多いことが分かる。欠落音については、赤とんぼでは同一音高の連続箇所が楽譜上4箇所存在しており、それらが1音のロングトーンに変換された影響が見られた。余分音が多い原因は歌唱中のピッチ変動や揺れが多いためである。例えば3小節目の「あか」のような落差の大きい箇所では、ピッチが大幅なアンダーシュートを起こし、本来の音高に戻るまでに複数の音高に掛かる。また3-4小節にかけての「とーんーぼー」のようなロングトーンは意図しないピッチ変動が起きやすい。またBP2では、タップの有無に関わらず同等の認識精度であり、タップを行うことによって歌唱が乱れて精度が下がるようなことは無かったと考えられる。

総じて、TVMはBP2よりも再現率・適合率・F値いずれも全被験者について高い結果を示した。再現率・適合率ともに100%の被験者が5名いた。これには楽器経験なしの被験者A、Bも含まれており、このレベルの曲や歌唱条件に対しては楽器経験の有無は影響を及ぼしにくいと見られる。

5.2 赤とんぼ:テンポ BPM = 120

「テンポ BPM = 120、歌詞歌唱、タップあり」の歌唱条件による入力3回分計93音について被験者ごとに集計を行った結果、およびBP2におけるタップの有無による精度比較用に「テンポ BPM = 120、歌詞歌唱、タップなし」の結果を表4-Bに示す。

TVMでは歌唱テンポの上昇に伴い負荷が高まるとともに誤り・欠落・余分の各音数も自由テンポ時より増加しているが、これは妥当な結果と言える。中でも被験者Eは欠落・余分音が大きく増加しているが、音長をある程度保ったタップ間隔ではなく、区切るべき箇所から全く外れた音の途中でタップされた例が見られたことから、テンポが速く追いつかなかったというよりもタップするべき位置を把握できずに混乱したと見られる。

一方BP2では余分音については、自由テンポ時よりもむしろ減少する結果となった。これは、テンポが速くなると1音当たりの歌唱時間が短くなりピッチの変動が減るためと考えられる。またBP2では、BPM=120での歌唱でも自由テンポ時と同様タップの有無によらず同等の認識精度であり、タップの有無はあまり精度に影響しなかったと考えられる。

総じて、タップ位置のミスが音高変換精度を落とす

のは TVM の性質上避けがたく、テンポ自由時よりは多少劣るものの、再現率・適合率・F 値いずれもほとんどの被験者について TVM の方が高い結果となり、再現率・適合率ともに 100%の被験者が 2 名いた。また余分音の出力が十分に抑制されており、テンポが速くなっても正しく変換可能であることが分かった。

5.3 自由曲

各被験者が選択した自由曲について「テンポ BPM = 自由、歌詞歌唱、タップあり」で入力した結果を図 4-C に示す。図 4-C に見られるとおり、合計値では TVM が BP2 よりも再現率・適合率・F 値のすべてにおいて上回り、総合的にみると TVM は、「タップしながら歌唱する」という負荷の高さにも関わらず、より実践的なポップスなどのメロディの入力においても高い変換精度を得られていることが分かる。

ただし、問題点も明らかになった。被験者 A, E, F については、欠落音中の結合音の数が多く見られる結果となっている。結合音は、被験者が 1 音ごとに正しくタップしていないため複数音が 1 音に結合されて変換された箇所であることを示す。TVM では区間の最頻音高が採用されるため、複数音が 1 音に結合された場合、最長音長の音の音高が採用されてしまい、その結果として誤り音と判定され、更に残りの音は欠落音と判定されてしまう。よって結合音の存在は誤り音と欠落音の両方に影響を与えてしまう結果となる。

ただし、今回の評価基準では、タップ開始時点の音の音高を正解として精度を評価しているが、仮に複数音が結合されて 1 音にみなされてしまった場合に、そこに含まれる音のいずれかの音と音高が一致した場合も正解とみなせば、精度は更に上がる。これは BP2 でも同様に起こるが、音高変化で音が区切られるため、結合音の発生は主に同一音高連続箇所となる。よって TVM のように 1 音目が一致しなくても他のいずれかの音が一致することによる精度向上の余地は少ないと言える。このように今回の評価基準は TVM にとって厳しいものであるにも関わらず、TVM では、被験者 E, F の場合に再現率についてそれぞれ BP2 より 15% および 18% 高く、被験者 A の場合に BP2 と同等の適合率であり、また被験者 F の場合に適合率が BP2 よりも 14% 高いという結果となっていることから、TVM は良好な性能を達成していると言える。

その他、A, E, F 以外の被験者における誤りの発生原因は、タップ開始位置のズレにより音区切りがうまくいかなかったことにあると考えられる。テンポが速く追いつかなかったと想像される箇所と、タップす

るべき位置を把握できずに混乱したと想像される箇所が、ともに存在した。しかしながら、各被験者とも非常に高いと思われる負荷にも関わらず高い再現率を達成していることから、「タップしながら歌唱する」行為は、基本的に実施可能なものであったと言えることができるだろう。

5.4 全体考察

以上より、TVM は、BP2 のような音高変化によって音を区切る VtoM システムの問題点である 1 音が複数音に認識され余分な音出力されやすいという点に対処できることが示された。また、TVM システムは、歌唱時の負荷の増加はあるものの、既存の歌詞歌唱などの任意の発音の歌唱を許容するシステムに比べて、より高い音高変換精度を達成した。よって先の「タタタ歌唱」システムとの比較結果 5) と合わせて、TVM は十分な有用性があると考えられる。

6. 結論

本稿では、我々が提案しているメロディリズムタップによって音の区切りを入力する人間と計算機との協調的 VtoM である、タップ併用 Voice-to-MIDI システムと歌詞歌唱などの任意の発音の歌唱を許容する既存 VtoM システムとの音高変換精度の比較を行った。その結果、TVM の有用性を実証するとともに、VtoM における音の区切りの重要性を示した。

今後、誤った音区切りを減らすことと、タップへの依存度を減らすために必要なタップか否かを判定する機構を開発し組み込む予定である。また歌詞先作曲における実践的な使用評価を行っていく予定である。

参考文献

- 1) YAMAHA 株式会社 : XGworks ST; http://www.yamahasyth.com/jp/products/music_production_software/ma_65w/
- 2) 株式会社 INTERNET: SingerSongWriter Lite5; <http://www.ssw.co.jp/products/ssw/win/sswlt60w/index.html>
- 3) MakeMusic Inc.: Finale2010, <http://www.e-frontier.co.jp/>
- 4) 伊藤 直樹, 西本 一志: 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, 情報処理学会研報 2006-EC-5, Vol.2006, pp.43-48, (2006).
- 5) Naoki Itou, Kazushi Nishimoto: A voice-to-MIDI system for singing melodies with lyrics, Proc. of the int. conf. on ACE'07, pp.183-189, Salzburg, Austria, (2007).
- 6) 番弘光, 伊藤克亘, 武田一哉, 板倉文忠: タッピングを利用した音声認識の検討; 情報処理学

- 会研報, SLP-47, pp71-76, (2003).
- 7) 岩田憲治, 渡邊康司, 中川竜太, 篠田浩一, 古井貞熙: 音声とペンの準同期入力に対するマルチモーダル認識; 日本音響学会 2006 年秋季講演論文集 1-2-23, (2006).
- 8) 尾関弘尚, 鎌田貴幸, 後藤真孝, 速水悟: 歌声の歌詞認識における音高の影響について; 日本音響学会秋季講演論集, pp637-638, (2003).
- 9) 清水 純, 丸山 剛志, 三浦 雅展 柳田 益造: ハミングによる単旋律の自動採譜; 日本音響学会音楽音響研究会研資, Vol.23, No.5, pp.95-100,

- (2004).
- 10) 野ばら社: <http://www.nobarasha.co.jp/>
- 11) 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).
- 12) 株式会社河合楽器製作所: Band Producer 2, <http://www.kawai.co.jp/>

表 4 赤とんぼおよび自由曲の変換結果

- 注1. "*"付きの被験者は「音楽経験なし」と回答した被験者。
 注2. 全歌唱音数は本来 93 音だが、歌唱されなかったり、音高の特定が困難等で集計から除外した箇所がある。
 注3. 全歌唱音数(音) = 正解音数 + 誤り音数 + 欠落音数
 注4. 再現率(%) = 正解音数 / 全歌唱音数 * 100
 適合率(%) = 正解音数 / (正解音数 + 誤り音数 + 余分音数) * 100
 F 値 = (2 * 再現率 * 適合率) / (再現率 + 適合率)
 注5. 太字: 3 歌唱条件中最も高い値, 下線: BP2 のタップあり/なしの 2 条件を比較し、より高い値を示す。

A) 赤とんぼ [歌唱条件: テンポ自由, 歌詞歌唱, タップあり]

被験者	全歌唱音数(音)	TVM							BP2							BP2(タップなし歌詞歌唱)							
		正解(音)	誤り(音)	欠落(音)	余分(音)	再現率(%)	適合率(%)	F値	正解(音)	誤り(音)	欠落(音)	余分(音)	再現率(%)	適合率(%)	F値	全歌唱音数(音)	正解(音)	誤り(音)	欠落(音)	余分(音)	再現率(%)	適合率(%)	F値
A*	93	93	0	0	0	100	100	100	87	0	6	14	<u>93.5</u>	86.1	89.7	93	85	0	8	10	91.4	<u>89.5</u>	<u>90.4</u>
B*	93	93	0	0	0	100	100	100	80	1	12	6	86.0	92.0	88.9	93	58	5	30	3	62.4	87.9	73.0
C*	92	88	4	0	0	95.7	95.7	95.7	73	1	18	4	79.3	<u>93.6</u>	85.9	87	81	0	6	9	<u>93.1</u>	<u>90.0</u>	<u>91.5</u>
D*	93	92	1	0	0	98.9	98.9	98.9	90	0	3	13	96.8	87.4	91.8	93	91	0	2	11	<u>97.8</u>	<u>89.2</u>	<u>93.3</u>
E	93	91	2	0	2	97.8	95.8	96.8	88	0	5	9	94.6	<u>90.7</u>	<u>92.6</u>	93	80	4	9	9	86.0	86.0	86.0
F	93	93	0	0	0	100	100	100	90	1	2	28	96.8	<u>75.6</u>	<u>84.9</u>	92	90	0	2	31	<u>97.8</u>	74.4	84.5
G	93	92	1	0	0	98.9	98.9	98.9	90	1	2	14	<u>96.8</u>	85.7	90.9	93	90	0	3	12	<u>96.8</u>	<u>88.2</u>	<u>92.3</u>
H	93	93	0	0	0	100	100	100	87	0	6	2	93.5	97.8	95.6	93	90	0	3	4	96.8	95.7	96.3
I	93	93	0	0	0	100	100	100	90	0	3	5	96.8	<u>94.7</u>	<u>95.7</u>	93	93	0	0	21	100	81.6	89.9
合計	836	828	8	0	2	99.0	98.8	98.9	775	4	57	95	92.7	<u>88.7</u>	<u>90.6</u>	830	758	9	63	110	91.3	86.4	88.8

B) 赤とんぼ [歌唱条件: テンポ BPM = 120, 歌詞歌唱, タップあり]

被験者	全歌唱音数(音)	TVM							BP2							BP2(タップなし歌詞歌唱)							
		正解(音)	誤り(音)	欠落(音)	余分(音)	再現率(%)	適合率(%)	F値	正解(音)	誤り(音)	欠落(音)	余分(音)	再現率(%)	適合率(%)	F値	全歌唱音数(音)	正解(音)	誤り(音)	欠落(音)	余分(音)	再現率(%)	適合率(%)	F値
A*	93	93	0	0	2	100	97.9	98.9	76	0	17	13	81.7	85.4	89.7	93	77	0	16	11	<u>82.8</u>	<u>87.5</u>	85.1
B*	93	93	0	0	3	100	96.9	98.4	76	0	17	5	<u>81.7</u>	93.8	87.4	93	76	1	16	3	<u>81.7</u>	<u>95.0</u>	<u>87.9</u>
C*	93	85	7	1	1	91.4	91.4	91.4	54	2	37	0	58.1	<u>96.4</u>	72.5	93	62	1	30	4	<u>66.7</u>	92.5	<u>77.5</u>
D*	93	93	0	0	0	100	100	100	88	2	3	7	94.6	90.7	<u>92.6</u>	93	79	0	14	4	84.9	<u>95.2</u>	89.8
E	93	73	5	15	11	78.5	82.0	80.2	62	1	30	6	66.7	89.9	76.5	92	69	1	22	8	75.0	88.5	81.2
F	93	90	3	0	2	96.8	96.8	96.8	67	0	26	3	72.0	95.7	<u>82.2</u>	93	63	0	30	2	67.7	96.9	79.7
G	93	90	1	2	2	96.8	96.8	96.8	80	2	11	11	<u>86.0</u>	<u>86.0</u>	<u>86.0</u>	93	80	0	13	14	86.0	85.1	85.6
H	93	93	0	0	0	100	100	100	71	0	22	1	76.3	<u>98.6</u>	<u>86.1</u>	92	72	0	20	4	<u>78.3</u>	94.7	85.7
I	93	92	1	0	0	98.9	98.9	98.9	83	0	10	3	89.2	96.5	92.7	93	82	0	11	4	88.2	95.3	91.6
合計	837	802	17	18	19	95.8	95.7	95.8	657	7	173	49	78.5	<u>92.1</u>	84.8	835	660	3	172	54	79.0	<u>92.1</u>	<u>85.1</u>

C) 自由曲 [歌唱条件: テンポ BPM = 自由, 歌詞歌唱, タップあり]

被験者	全歌唱音数(音)	TVM							BP2							
		正解(音)	誤り(音)	欠落(音)	余分(音)	再現率(%)	適合率(%)	F値	正解(音)	誤り(音)	欠落(音)	余分(音)	再現率(%)	適合率(%)	F値	
A*	120	87	13	20	0	72.5	87.0	79.1	93	4	23	7	9	77.5	87.7	82.3
B*	63	58	5	0	0	92.1	92.1	92.1	44	1	18	7	2	69.8	93.6	80.0
C*	61	51	10	0	0	83.6	83.6	83.6	17	4	40	14	0	27.9	81.0	41.5
D*	122	121	1	0	0	99.2	99.2	99.2	99	0	23	16	20	81.1	83.2	82.2
E	98	80	10	8	10	81.6	80.0	80.8	65	0	33	10	4	66.3	94.2	77.8
F	172	155	8	9	2	90.1	93.9	92.0	124	1	37	21	31	72.1	79.5	75.6
G	90	90	0	0	0	100	100	100	66	1	23	14	12	73.3	83.5	78
H	198	193	3	2	0	97.5	98.5	98.0	141	1	57	43	0	71.2	99.3	82.9
I	209	197	12	0	1	94.3	93.8	94.0	166	2	41	17	7	79.4	94.9	86.5
合計	1133	1032	62	39/34	13	91.1	93.2	92.1	815	14	295/149	85	71.9	89.2	79.6	