論　文

Voice-to-MIDIのためのメロディリズムタップを用いた音数・音高の判定手法の提案

伊藤　直樹†*　　西本　一志††

A Method of Note Counting and Pitch Extraction by Using Melody Rhythm Taps for Voice-to-MIDI System

Naoki ITOU†* and Kazushi NISHIMOTO††

あらまし　計算機を用いた音楽制作における MIDI ファイル作成法の一つに鼻歌入力法（Voice-to-MIDI）がある．しかし既存システムでは 1 音ごとの区切りがうまくいかないことによって，出力された MIDI ファイルに欠落音や余剰音の発生，音高の誤判定などの変換精度低下が起こる．この問題に対して，幾つかのシステムでは，歌詞を全て「タ」に置き換える「タタタ歌唱」をさせることで音区切りの精度向上を図っている．しかし，歌詞先作曲のように歌詞歌唱によって，歌詞のイントネーションをメロディづくりに活用したい場合には不向きである．そこで我々は，Voice-to-MIDI の音数・音高判定精度の向上のために，歌唱と同時にタップをすることによってメロディリズムの区切りを入力する，人間と計算機の協調的な音数・音高判定手法を提案する．本手法と，タタタ歌唱を前提としない，自由歌唱可能な既存システム 3 種類を比較した結果，欠落する音や不要な音の発生が抑制され，音数及び音高判定精度が向上することを確認した．また，楽器経験の有無がタップに影響しないこと，そしてタップの有無は歌唱に影響しないことを示す．

キーワード　音区切り，音高判定，鼻歌入力，歌詞歌唱，歌唱同期タップ

1. まえがき

計算機を用いた音楽制作における MIDI（Musical Instrument Digital Interface）ファイル作成法の一つに，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，例えば，カラオケ等の歌唱は得意だが絶対音感や相対音感をもたないユーザや，多くの音楽編集ソフトで楽譜データの入力手段として採用されているリアルタイム入力を楽器演奏技術がないためにできないユーザを支援できる入力方法である．また，楽器演奏技術があるユーザにとっても，例えばキーボードパートの入力はキーボードで，ボーカルパートの入力は VtoM で，といったパートに即した入力方法の使い分けなどのメリットがある．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，おおよそ

（1）　歌唱区間の検出

（2）　1 音ごとの区間検出

（3）　その区間内で短時間 F0 推定を繰り返し，当該区間全体にわたる短時間 F0 の集合を取得

（4）　その F0 推定情報からの区間音高判定

（5）　得られた音高・音長から音符列を作成

という処理段階に分類できる（(1) が明確に存在しなかったり，(2) の区間検出と (3) の短時間 F0 推定と短時間 F0 集合取得の処理順序が前後したりするなど，全てのシステムがこのとおりとは限らない）．

この各段階で得られた結果は，いずれも連鎖的に次の処理の結果に影響を与える．例えば，(2) の処理で誤った区間が検出されると，音数が変化するのみなら

† 北陸先端科学技術大学院大学知識科学研究科，能美市
School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1–1 Asahidai, Nomi-shi, 923–1292 Japan
†† 北陸先端科学技術大学院大学ライフスタイルデザイン研究センター，能美市
Research Center for Innovative Lifestyle Design, Japan Advanced Institute of Science and Technology, 1–1 Asahidai, Nomi-shi, 923–1292 Japan
* 現在，インターメディアプランニング株式会社に所属

ず，(3) の処理で区間内での短時間 F0 の分布も変化し，結果として (4) の処理で誤った区間音高判定が行われてしまう．したがって初期の段階での誤りは，それ以降の段階の誤りにもつながり，最終的に得られる音数や音高の変換結果を極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段階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，歌唱区間の検知及び 1 音ごとの区間検知の精度を上げることは，それ以降の処理段階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 音ごとの区間を計算機処理によって検知することは容易とはいえない．このため，幾つかの VtoM システムでは,「タタタ～タタ」のように全ての歌詞を「タ」に置き換えて明確に区切る「タタタ歌唱」のような，特殊な歌唱方法が求められる．これにより一定水準の処理結果が得られるようになる．しかしながら，例えば先に歌詞を作ってからメロディを作曲する「歌詞先作曲」[4], [5] の場合，歌詞のもつイントネーション等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．

また，歌唱などの有声区間全体ではなく，任意位置・区間のみを切り出して処理する用途（音楽療法支援に関する研究 [6] など）では，例えば処理させたい位置だけ「タタタ歌唱」するのは難しく，歌唱や発声としても不自然である．特に文献 [6] の事例では，発声者が認知症患者であるため，発声者自身が切り出しを行うことができず，かつ発声者に歌唱や発声について指示を与えることも難しい．このような場合，タタタ歌唱を前提とするようなシステムは適用できない．

以上で示したような場合では，歌唱スタイルを制限せず，任意のスタイルの歌唱を許容できる VtoM システムの実現が求められる．

そこで我々は，これらに対応する Voice-to-MIDI 変換手法の実現に向けて，人間が歌唱や発声などに合わせて 1 音の区切りをタップ入力する，計算機との協調的な音数・音高判定手法を提案する．

これによって，人間が自らの歌唱と同時にタップでメロディリズムを入力し，1 音の区間をより明確に設定することによって音数と音高の判定精度を高めたり，他人の歌唱や発声に対してタップによる音区切りを行い，任意位置・区間の音高を取得することが可能になる．本論文では，これらの用途のうち，自らの歌唱と同時にメロディリズムをタップする用途を対象に，提案手法の実装システム（タップ併用型 Voice-to-MIDI：TVM と略す）と，歌詞歌唱などの任意発音の歌唱を許容する既存 VtoM システムとで音数・音高の変換精度を比較する．また，楽器経験の有無のタップへの影響やタップの有無の歌唱への影響の評価を行う．その結果，タップの付加により音数の抽出の正確さが増し，それが音高判定の精度向上にも寄与すること等を示す．

以下，**2.** では関連研究について概観し，本研究の位置づけを行う．**3.** では提案手法の詳細と，これに基づいて構築したプロトタイプシステムの構成について述べる．**4.** では，提案手法と既存システムとの比較実験を示し，**5.** でその結果及び提案手法の有用性と課題について議論する．**6.** はまとめである．

2. 先行研究

VtoM システム [7]～[10] や VtoM を応用した Query By Humming（QBH）と呼ばれる楽曲検索インタフェース [11]～[14] は多数存在する．音の区切りに時間軸上で歌唱のパワーが大きくなる箇所の検知や F0 の遷移等を用いるものが多く，いずれも歌唱とは別に音を区切る情報を入力するものではない．

歌唱や発声以外の情報も使って入力を行っている研究として，文献 [15], [16] 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．しかし，VtoM の用途には各区間の音高判定処理が必要となる．また，商品 [17] に搭載されている Step Entry モードでは，声から音高を取得する間に音価をマウスで入力可能である．熟達者であれば，歌唱と同時に音価を入力できることもありうるが，本質的にはステップ入力であり，リズムや音価を理解しておく必要がある．

人間と計算機が協調して採譜するシステムとして，半田らは発音時刻の候補を音楽情景分析器で求めて表示し，人間が音の有無や音高の上行・下行の情報を入力するシステム [18] を提案した．しかし，視覚情報による協調であり，聴覚情報を用いた本研究とは協調の方法が異なる．

3. タップ併用型 Voice-to-MIDI システム

本章では，協調的な音数・音高判定を実現する手法の評価のために，それらを実装したタップ併用型 Voice-to-MIDI システムについて述べる．

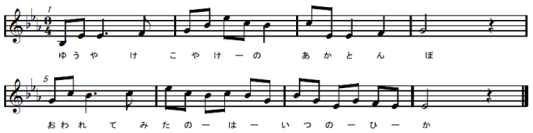

図 1　赤とんぼの楽譜作曲：山田耕作，作詞：三木露風
Fig. 1　Score of “Aka tombo”: compositon by Kosaku Yamada, lyric by Rofu Miki.

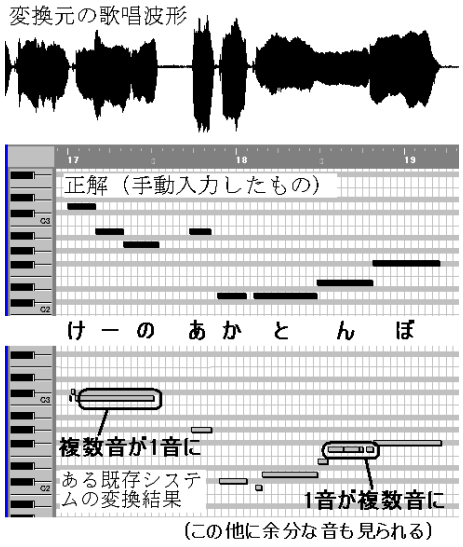

図 2　音量によって区切られたと推測される，複数音が 1 音に，1 音が複数音に変換された例（赤とんぼの「けーのあかとんぼ」）
Fig. 2　Samples of segmentation mistake with note binding and divorcing.

3.1　既存 VtoM システムの問題点

最初に既存の VtoM 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 VtoM システムに童謡「赤とんぼ」（野ばら社刊「童謡」の変ホ長調版 [19] 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）けーのあかとんぼ」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，中段は音区切りの比較のために正解のメロディラインを手動入力したもの（正解データ），下段はシステムによる認識結果をピアノロールで示す．このシステムは主に音量変化で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，逆に複数音存在する箇所が 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

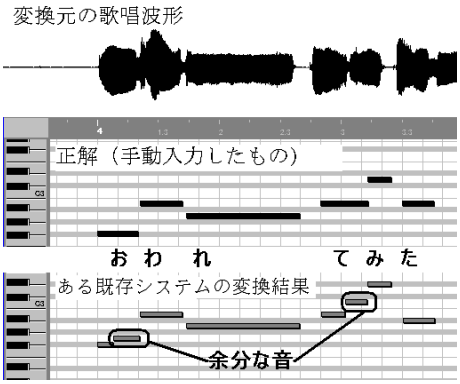

図 3　音高変化によって区切られたと推測される，余分な音が出力された例（赤とんぼの「おわれてみた」）
Fig. 3　Samples of segmentation mistake with extra notes.

よって音が区切られると推測されるが，意図しない音高変化にも反応してしまい，「お」と「て」の部分で余計な音が出力されてしまっている．

このように，従来の VtoM 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いえる．

総じて，以下のような箇所や条件において区切りミスが見られた．

- 同一音高の連続
- 激しい音量変化
- 大きい音高変動
- 不十分な音高変動
- 歌詞（任意発音）歌唱
- 環境音の誤入力

3.2　タップ併用型 VtoM（TVM）手法の概要

上記のような問題に対処するためには，音量変化が乏しくて音が区切られない問題や音高変化などによる意図しない区切れの発生の抑止，不要区間の除去が必要となる．そこで TVM では，計算機が苦手とするが人にとっては容易な区間区切りを人が担当し，計算機は得意だが人が苦手としやすい F0 推定を計算機が担当する，人と計算機の協調型の処理機構を採用した．

具体的には，ユーザは，歌唱するメロディのリズムに併せて鍵盤楽器や PC キーボードなどのデバイスをタッピングし，メロディの各音を区切る情報（リズム区切り情報）を入力していく．一方システムはリアルタイム処理で，歌唱から音高，リズム区切り情報からリズムと音長を取得し，最終的にマージして出力する．

3.3 プロトタイプの構成

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 = 440 Hzを基準とする）をもったMIDIデータである．入力音声は22050 Hz，16 bit，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キーボードの打鍵及び離鍵の入力時刻情報を用いる．PCキーボードの場合は，タップに「,」及び「.」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．以下に1音ごとの区間検知と，各区間における音高判定の処理手順を示す．

（1） キーが押下され，システムに押鍵情報が入力されたら，これをトリガとしてマイクより入力される歌唱音声データに対して，後述するF0推定処理を開始する．

（2） キーが離されたら，その離鍵情報が入力された時点か，歌唱の途切れが検知された時点（これは後述する無発声検知機構によって決定される）の，いずれか時間的に後の方が1音の区間の終了となる．タップ開始から区間の終了までを音長として，その区間内でF0推定処理を繰り返す．

（3） 1音の区間終了後，F0時系列データから半音単位のヒストグラムを生成し，最頻音高の音名を求め，これをこの区間の音高として出力する．

F0推定は，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=2048 samples：約100 ms，フレーム移動間隔=128 samples：約6 ms)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る．更にスペクトルの内挿[20]を用いてcent単位で音高推定してF0推定結果として出力する．これは周波数解像度不足を補うためである．

本システムでは，タップ開始時刻について，区切り情報と波形の同期が必要となる．PCキーボードのキーを叩いたときのKeypressイベントの時刻と打鍵音（パルス音）の録音時刻とのずれを調査したところ，試作システムでは，おおむね1024 sample（約50 ms）分Keypressよりも遅れて録音されたため，1024 sample分調整して同期精度を高めた．

3.4 無発声検知機構

予備実験において被験者のタップ方法を観察したところおおむね2通りとなった．一つは，1音の歌唱終了までキーを押下し続けるタップであり（図4のタップ法1），もう一つは，押下してすぐ離してしまうようなタップである（図4のタップ法2）．

タップ法1のみに対応したシステムでは，タップした時間がそのまま音長になるため，タップ法2が行われたときに音長が極端に短くなったり，十分な量のF0推定情報が取得できなくなる問題が見られた[21], [22]．そこで，歌唱区間の途切れを検知する機構によって，たとえタップが早期に終わってもそこで歌唱終了とみなされないようにした．

具体的には，循環自己相関の結果，タップ終了後でもD2-F5の音高範囲内に最大の正相関値が存在する限りフレーム移動間隔約6 ms分区間が順次延長され，なくなれば歌唱の終了と判断するようにした．

この機構により，音長は，タップ終了と歌唱終了のタイミングで以下の3パターンに定められる．

（1） タップ終了後に歌唱終了：歌唱終了時点

（2） 歌唱終了後にタップ終了：タップ終了時点

（3） 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

ただし，タップ開始から200 ms未満までは遅れて歌唱開始されても歌唱終了を誤って検知されないようにした．タップ開始時に歌唱がない場合，即座に歌唱が終了したとシステムが誤検知してしまうと，パターン（2）が適用されて，歌唱の有無にかかわらず，必ずタップ終了時点までが1区間になってしまう．これを防ぐためである．

この値は，第一著者がどれぐらいまで自然に歌唱とタップをずらし得るかを実験で調査して経験的に得た値を基に，システムに慣れないユーザを考慮して余裕をもたせた値である．

また，F0推定が上手くいかず，音があるのに音高範囲内にF0がな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FFT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．

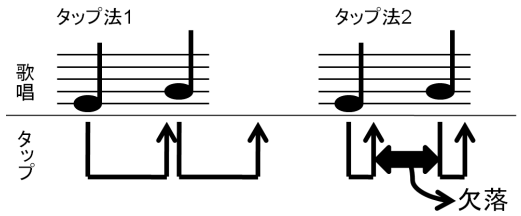

図4 2種類のタップ方法
Fig. 4 2 types of tapping manner.

この無発声検知機構によって，対象とする音高範囲内に他に目立つ音がなければ，音量しきい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．しかし，我々が実使用において想定しているマイクである，比較的感度が低い PC 内蔵マイクやヘッドセットマイクなどのマイクで調査したところ，歌唱終了と推定できる位置から大きく外れることなく 1 音の区間が終了した．

3.5 プロトタイプシステムの仕様的限界

プロトタイプシステムが仕様として対応できる音域及びテンポ（タップ速度）の限界について述べる．

音域については，ポップス楽曲を想定し，A4 = 440 Hz を基準として，下限を D2，上限を F5 とした．これは，おおよそバス歌手～アルト歌手の音域に相当する（文献 [23]）．メゾソプラノやソプラノの音域には対応していないが，ポップス等でよく使われる音域に対しては十分と考える．

テンポについては，FFT フレーム移動間隔が約 6 ms なので，この間隔を 16 分音符とし，人間が 6 ms ごとにタップできると仮定すれば，無発声検知機構の「歌唱が終了しないまま次のタップ開始」のパターンによって原理的には BPM=2500 程度まで対応できる．しかし実際の入力では，それほど早く歌唱やタップをすることはなく，BPM=250 程度まででよいと思われるため，本プロトタイプシステムは十分対応している．

4. 評価実験

4.1 実験概要

提案手法の検証のため，前章で述べた TVM プロトタイプシステムを用いて，歌唱音声に対する音区切り（音数）と各区間の音高判定精度を評価するとともに，楽器経験のタップへの影響及びタップの有無の歌唱への影響を調査した．

なお，この実験の評価対象は，システム自体の性能であり，入力者の歌唱やタップの技術に依存する内容については評価の対象とせず，極力排除した．例えば，歌が下手で楽譜どおりの変換結果にならなかったとしても，それだけではシステム自体の性能の善しあしはいえない．この場合，楽譜どおりの歌唱かどうかではなく，実際の歌唱の音高を割り出し，それとシステムの変換結果との比較を行うことによってシステム自体の性能の善しあしが分かる．

また，システムの仕様として対応可能な音域やテンポ（タップ速度）の限界については **3.5**「プロトタイプシステムの仕様的限界」に記した．

評価では，TVM と同様に歌詞歌唱などの自由な発音による入力を許容し，歌唱スタイルを制限しない入力という我々の目的により近いと思われるシステムを比較に用いた．

評価項目は以下とした．

（1） 任意発音歌唱に対する性能
（2） 歌唱同期タップの実施可能性
（3） (2) における楽器経験の影響
（4） タップの歌唱への影響

(1) と (2) については，後述する **5.1**～**5.3** で曲及び歌唱条件ごとに評価し，(3) は **5.4** で TVM の結果を用いて楽器経験の影響について評価する．(4) については **5.5** で比較 3 システムのタップあり歌唱の処理結果とタップなし歌唱の処理結果とを比較する．

4.2 楽　　曲

歌唱する楽曲は以下の 2 種類である．

（1） 課題曲（赤とんぼ）
（2） 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，適度に難しい．そしてよく知られ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，速いテンポでも歌唱とタップの同期が可能かを検証した．

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 比較に用いた VtoM システム

比較に用いた VtoM システムは，3 種類である．

（1） CMP：音高変化に基づいて区切る先行研究システム
（2） RYN：先行研究のシステム [10]
（3） BP2：商用で市販されているシステム [25]

CMP は，第一筆者がこの実験を行うにあたって区

切りの手動・自動の比較のために作成した．F0 推定法などは TVM と同様とし，タップによる区切りの代わりに音高の変化で区切る．音高を区切る基準については，文献 [24] を参考に 50 cent 以上の差があるときとした．無発声検知機構は，判定精度が低下したので実装しなかった．また，約 70 ms 以上の音長のみ変換するようにした．これは予備調査により，速いテンポへの対応，できるだけ多い認識音数，不要な音の誤変換の少なさのバランスを考慮した値である．16 分音符換算で BPM=213 程度までの歌唱テンポに対応可能である．

RYN は，先行研究との比較のため用いた．文献 [10] の筆者らからシステムの Linux バイナリの提供を受け，そのまま使用した．これは楽曲中からメロディーライン等を抽出し，MIDI データへの変換を行うシステムであり，文献 [9] 等，Ryynanen らが保有する技術を応用して構築されたシステムである．音の区切りは，“Accent Signal” と呼ばれる FFT フレーム中のスペクトルエネルギーの量を用いて行っている．

BP2 は，KAWAI: Band Producer 2 に付属の鼻歌入力機能である．この機能は，あらかじめ設定した音量しきい値を超過したときと半音単位の音高しきい値を超えたときに音符が区切られる仕様であると，変換結果から推測される．音高変化があれば区切られるため，歌詞歌唱にも対応していると考えられる．

4.4 機材設定

TVM においてタップに用いたデバイスは，HP: 2710p ノート PC のキー「,」及び「.」である．これらのキーは隣接して存在し，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．また，歌唱収録用マイクは Shure: SM87A を用いた．

次に各種情報の記録及び処理手順について述べる．2 台の PC を用意し，PC1 では，被験者に試唱させて BP2 の録音音量しきい値を設定した後，BP2 に伴奏なし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 上で記録される．

PC2（2710p ノート PC）では，TVM のために，歌唱と同時に行ったタップ区切りの情報を自作ソフトで記録する．このタップ情報と PC1（BP2）で記録した波形とを組み合わせてバッチ処理で MIDI データに変換する．実験では全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，本来リアルタイム処理である TVM をバッチ処理とした．

また，PC1 で記録した歌唱波形と TVM のタップ情報の同期が必要となるが，PC2 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 PC1 の波形を目視して同期位置を探した．具体的には，PC1 と PC2 の両波形に共通する特徴的な形状の箇所を複数探し，それらの箇所の間隔が両波形で一致するかを評価して同期位置を決定した．なお相互相関などで自動同期推定を行っても，最終的に目視による確認が必要であると考えて自動処理は行わなかった．CMP と RYN は，いずれも BP2 で取得した波形を，必要があれば Adobe: Audition 1.0 で対応サンプリングフォーマットに変換した後，バッチ処理した．

4.5 被験者

被験者は，筆者らが所属する大学院の男子学生 8 名と女子学生 1 名である．TVM の支援対象は，主に音感をもたないユーザであるが，実験では様々なデータを得るために和音楽器やリズム楽器の経験者，音感があると思われる学生にも参加をお願いした．

どのような被験者が参加したかの傾向を知るために，予備調査により被験者の音楽知識や能力，楽器経験を調べた．項目を以下に示す．

（1）「鍵の音名」：ピアノ上で指差された鍵を見て音名を回答

（2）「音高聴取」：ピアノで弾かれた単音の音名を回答

（3）「音の高低」：ピアノで弾かれた 2 音の高低を回答

各項目はいずれも全 6 問ある．「鍵の音名」では基礎的知識，「音の高低」では基礎的な知覚能力，「音高聴取」では高度な学習経験・技能を調査した．実験では，被験者は最低限歌唱が可能であればよく（タップは，全くできないようなレベルでなければ問題ない），被験者 9 名が歌唱に問題がないことは確認している．

これらの結果より，楽器経験なし 4 名と経験あり 5 名に分類した．各被験者の正解数と楽器経験を表 1 に示す．表 1 より，安定した歌唱が可能と考えられる「音高聴取」の成績が良い被験者がいる一方で，VtoM の支援対象となりうる，基礎的な「鍵の音名」や「音の高低」の正解数が少ない比較的音楽に詳しくない被験者も含まれており，経験の有無だけでは測れない様々なレベルの被験者がいることが分かる．

4.6 実験手順

実験は大学院内の防音室を用いて 1 名ずつ行った．まず VtoM の練習及び歌唱しながらタッピングする練

表 1　各被験者の予備調査項目（1）～（3）の正解数と楽器経験
Table 1　Results of pre-test and experiences of musical performing for each subject.

被験者	(1) 鍵の音名	(2) 音高聴取（正解）	(2) 音高聴取（半音差）	(3) 音の高低	楽器経験
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

注 1．被験者 A～D は「楽器経験なし」と回答した被験者
注 2．予備調査項目（2）「音高聴取」は，正解した個数と被験者が正解より半音ずらして判定した個数を示す．

表 2　各曲の歌唱条件
Table 2　Singing conditions for each song.

（A）赤とんぼ

テンポ	タップ
自由	あり
	なし
BPM = 120	あり
	なし

（B）自由曲

テンポ	タップ
自由	あり

習を 5 分ずつ行った後，以下の順序で実施した．最初に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．各曲の歌唱条件を表 2 に示す．課題曲ではタップありなしをランダムな順番で指示して歌唱させた．赤とんぼについては，それぞれ 3 回ずつ歌唱を入力させた．「BPM=120」で歌唱する場合は，メトロノームに合わせて歌唱するよう依頼した．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．また，全ての歌唱は無伴奏で行った．

表 3　各被験者の自由曲
Table 3　List of subject own-selected songs.

被験者	歌手名	曲名
A	Mr. Children	Over
B	井上あずみ	さんぽ
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

4.7 評価方法

被験者が必ずしも楽譜どおり，あるいはそれを移調した音高どお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判定性能を評価するために，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から正解の音高データを作成した．BP2 で記録した実験中の歌唱音響波形から，第一筆者(注1)が 1 音ごとに音高の特定を行った．また，正解の音高データと各システムの出力結果との時間同期や欠落音などの判定のために発音開始時刻と終了時刻の特定も同時に行った．これらを「正解データ」とした．作成された音列は必ずしも楽譜どおりの音高列とはならないが，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．

歌唱からの音高及び発音開始時刻と終了時刻の特定の方法（正解データの求め方）は以下のとおりである．

（1）　各音のおおよその区切りを試聴や波形の目測で割り出し，発音開始時刻及び終了時刻とする．

（2）　波形編集ソフト（Adobe: Audition1.0）上で各音の発音開始～終了までをループ再生させながら，ピッチベンドホイールつきのキーボード（Ensoniq: MR-76）を同時発音してうなりを聴き，音高特定を試みる．

（3）　1 音中で音高変化がある場合は，2～4 箇所程度の区間に分けて（歌い始め直後と歌い終わり付近は除く），局所的に音高特定を行う．

（4）　適宜波形編集ソフト上で目視計測した 1 波長の時間から周波数を逆算して用いた．

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．この作業により各音を，

（注1）：高校時代に男性合唱部に 3 年間所属した経験があり，また単音の音高を判定できる程度の絶対音感を保有している．

（1） 音高が一意に決まる音
（2） 2 音高の間で決めがたい音
（3） 分類（2）よりも明確に音高が変化する音

の 3 種類に分類した．また，(2) と (3) に分類される音は，可能性のある音全てを正解データとみなした．正解音高は 1 音につき 1 音高に定まるのが最良だが，音高のゆれが大きい場合など，1 音中でどの音高が優勢であるかを割り出すのは困難であるため，候補全てを正解とした．

なお，2 音から生じるうなりがなくなる周波数は客観的に一意に決まるため，作業者の違いによる正解データの大きな違いは生じにくいと考えられ，よって作業者が 1 名であることは妥当性を有すると考える．

次に個々の音について正解データと認識結果とを対応づけ，両者の音高を比較して正否を判定した．分類 (2), (3) に該当する音との比較では，複数ある正解データのうちいずれかの音高と一致すれば正解とした．最終的に表 4 のように分類された．

「結合音」とは，正しく区切られずに前後の音と結合した音を意味する．結合音の区間に一致する正解音列と比較したとき，先頭の音と結合音の音高が一致すれば結合音は「正解音」，不一致ならば「結合音による誤り音」に分類される．そして，残りの音は「結合音による欠落音」となる．

「誤り音」は，誤り音の全体数と，結合音によって生じた誤り音数に分けて示す．誤り音の全数と結合音による誤り音の差分は，F0 推定のミスによる誤り音数と考えてよい．

「欠落音」は，出力されなかった音の全体数と，結合音によって生じた欠落音数に分けて示す．これらの音数の差分は，そもそもシステムが認識しなかった音数となる．

「余分音」は，本来 1 音だが複数音に認識されたとき，必要な 1 音分を除いた残りの音，そして歌唱中における咳などのノイズである．1 音分については，複数音のいずれかの音が正解と一致すれば正解音，全くなければ誤り音に加算される．

表 4 認識結果の分類
Table 4 Categories for melody extracts.

カテゴリー	サブカテゴリー	説明
正解音	—	正解と一致した音
誤り音	—	正解と一致しなかった音
	全数	誤り音の全体数
	結合音による誤り音	他の音との結合で生じた誤り音
欠落音	—	欠落した音
	全数	欠落音の全体数
	結合音による欠落音	他の音との結合で生じた欠落音数
余分音	—	余分な音

各メロディの全歌唱音数（赤とんぼの場合正しく歌唱されれば 31 音）は，以下の式のように (1)〜(3) の合計で求まる．

全歌唱音数（音）= 正解音数+誤り音数+欠落音数

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は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，及びシステムが認識した全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適合率の二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．

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

5. 評価実験結果及び考察

評価実験結果及び考察について述べる．

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を表 5 に示す．

いずれの被験者とも TVM が最もよい再現率・適合率・F 値であった．5 名が再現率・適合率共に 100%であり，欠落音・余分音が十分抑制されていることが分かる．誤り音については，全て F0 推定ミスが原因であった．過不足のないタップによって欠落音・余分音が共に抑制され，タップ位置の大きなズレによる誤り音の発生もほとんど見られなかったことから，音数の切り出しや音高の判定に必要な歌唱同期タップができているといえる．

CMP・RYN・BP2 のいずれも正解音数自体は比較的多いが，TVM より欠落音が多く，また，欠落音中の結合音が，CMP（95 音中 58 音）と RYN（42 音中 23 音）では半数以上を占めた．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，それらがロングトーンに誤変換されやすいことが影響したと見ら

表 5　赤とんぼの変換結果　[歌唱条件：テンポ自由，歌詞歌唱，タップあり]
Table 5　Results of “Aka tombo”: [sung with own tempo, lyrics and taps].

被験者	全歌唱音(音)	正解(音)				上:誤り(音)/下:結合(音)				上:欠落(音)/下:結合(音)				余分(音)				再現率(%)				適合率(%)				F値			
		TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2
A*	93	93	76	78	87	0	5	12	0	0	12	3	6	0	13	29	14	100	81.7	83.9	93.5	100	80.9	65.5	86.1	100	81.3	73.6	89.7
						0	5	0	0	0	10	3	1																
B*	93	93	82	77	80	0	3	3	1	0	8	13	12	0	7	8	6	100	88.2	82.8	86.0	100	89.1	87.5	92.0	100	88.6	85.1	88.9
						0	2	1	0	0	8	3	2																
C*	92	88	81	86	73	4	1	2	1	0	10	4	18	0	18	14	4	95.7	88.0	93.5	79.3	95.7	81.0	84.3	93.6	95.7	84.4	88.7	85.9
						0	1	0	0	0	4	1	2																
D*	93	92	75	88	90	1	4	3	0	0	14	2	3	0	6	5	13	98.9	80.6	94.6	96.8	98.9	88.2	91.7	87.4	98.9	84.3	93.1	91.8
						0	4	0	0	0	14	2	3																
E	93	91	75	51	88	2	4	38	0	0	14	4	5	2	18	11	9	97.8	80.6	54.8	94.6	95.8	77.3	51.0	90.7	96.8	78.9	52.8	92.6
						0	0	0	0	0	2	4	2																
F	93	93	88	88	90	0	0	2	1	0	5	3	2	0	29	41	28	100	94.6	94.6	96.8	100	75.2	67.2	75.6	100	83.8	78.6	84.9
						0	0	0	0	0	4	3	2																
G	93	92	84	87	90	1	0	3	1	0	9	3	2	0	11	6	14	98.9	90.3	93.5	96.8	98.9	88.4	90.6	85.7	98.9	89.4	92.1	90.9
						0	0	0	0	0	4	2	2																
H	93	93	77	88	87	0	3	0	0	0	13	5	6	0	2	4	2	100	82.8	94.6	93.5	100	93.9	95.7	97.8	100	88.0	95.1	95.6
						0	3	0	0	0	5	2	3																
I	93	93	78	86	90	0	5	2	0	0	10	5	3	0	7	5	5	100	83.9	92.5	96.8	100	86.7	92.5	94.7	100	85.2	92.5	95.7
						0	5	0	0	0	7	3	2																
合計	836	828	716	729	775	8	25	65	4	0	95	42	57	2	111	123	95	99.0	85.6	87.2	92.7	98.8	84.0	79.5	88.7	98.9	84.8	83.2	90.6
						0	20	1	0	0	58	23	19																

注 1. “*” 付きの被験者は「楽器経験なし」と回答した被験者（表 6 も同様）
注 2. 欠落音の下段は欠落音中の結合音数，誤り音の下段は誤り音中の結合音に起因する誤り音数を示す．また，誤り音と結合音由来の誤り音の差分は F0 推定ミス由来の誤り音数を示す．(表 6 も同様)
注 3. 黒地白文字：タップあり歌唱で 4 システム中最もよい値を示す．ただし誤り・欠落音の下段の結合音と結合音由来の誤り音は対象外とする．(表 6 も同様)

表 6　赤とんぼの変換結果　[歌唱条件：BPM=120，歌詞歌唱，タップあり]
Table 6　Results of “Aka tombo”: [sung with BPM=120, lyrics and taps].

被験者	全歌唱音(音)	正解(音)				上:誤り(音)/下:結合(音)				上:欠落(音)/下:結合(音)				余分(音)				再現率(%)				適合率(%)				F値			
		TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2
A*	93	93	71	80	76	0	5	10	0	0	17	3	17	2	8	16	13	100	76.3	86.0	81.7	97.9	84.5	75.5	85.4	98.9	80.2	80.4	83.5
						0	4	0	0	0	8	3	3																
B*	93	93	76	77	76	0	3	2	0	0	14	14	17	3	2	11	5	100	81.7	82.8	81.7	96.9	93.8	85.6	93.8	98.4	87.4	84.2	87.4
						0	3	0	0	0	13	3	3																
C*	93	85	83	78	54	7	2	10	2	1	8	5	37	1	20	9	0	91.4	89.2	83.9	58.1	91.4	79.0	80.4	96.4	91.4	83.8	82.1	72.5
						0	0	0	0	0	3	1	1																
D*	93	93	67	75	88	0	9	16	2	0	17	2	3	0	3	1	7	100	72.0	80.6	94.6	100	84.8	81.5	90.7	100	77.9	81.1	92.6
						0	8	0	0	0	16	2	2																
E	93	73	79	77	62	5	1	11	1	15	13	5	30	11	17	3	6	78.5	84.9	82.8	66.7	82.0	81.4	84.6	89.9	80.2	83.2	83.7	76.5
						1	0	0	0	1	3	3	1																
F	93	90	81	80	67	3	3	5	0	0	9	8	26	0	15	18	3	96.8	87.1	86.0	72.0	96.8	81.8	77.7	95.7	96.8	84.4	81.6	82.2
						0	1	1	0	0	6	4	2																
G	93	90	77	88	80	1	2	3	2	2	14	2	11	2	7	6	11	96.8	82.8	94.6	86.0	96.8	89.5	90.7	86.0	96.8	86.0	92.6	86.0
						0	0	0	0	0	3	2	2																
H	93	93	81	89	71	0	2	0	0	0	10	4	22	0	3	3	1	100	87.1	95.7	76.3	100	94.2	96.7	98.6	100	90.5	96.2	86.1
						0	2	0	0	0	4	1	3																
I	93	92	86	88	83	1	2	2	0	0	5	3	10	0	2	6	3	98.9	92.5	94.6	89.2	98.9	95.6	91.7	96.5	98.9	94.0	93.1	92.7
						0	2	0	0	0	5	2	3																
合計	837	802	701	732	657	17	29	59	7	18	107	46	173	19	77	73	49	95.8	83.8	87.5	78.5	95.7	86.9	84.7	92.1	95.8	85.3	86.1	84.8
						1	20	1	0	1	61	21	20																

れる．

CMP・RYN・BP2 は，余分音も多かった．余分音が多い原因は歌唱中の音高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，音高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音高変動が起きやすい．

総じて，TVM は欠落音や余分音等の問題を解決し，任意発音歌唱に対して高い性能を実現可能といえる．

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を表 6 に示す．

全体傾向としては，自由テンポ時よりも正解音数が減少が見られる．変化がないように見える RYN についても，正解音数に極端に差がある被験者 E を除くと減少している．

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いえる．中でも被験者 E は欠落音・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．しかし，全体では比較 3 システムよりも欠落音・余分音が十分に抑制されており，テンポが速

くなっても音の切り出しや音高判定に必要なタップが可能な被験者が多いことが分かった．

比較 3 システムについては，余分音が自由テンポ時よりも減少している点が特徴として挙げられる．これは，テンポが速くなると 1 音当りの歌唱時間が短くなり音高変動が減るためと考えられる．

総じて，タップ位置のミスが音高判定精度を落とすのは 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 が高い結果となり，特に 2 名において再現率・適合率共に 100%であったことから任意発音歌唱に対して性能が向上したといえる．

5.3 自 由 曲

各被験者が選択した自由曲について「テンポ自由，歌詞歌唱，タップあり」で入力した結果を表 7 に示す．表 7 より，合計値では TVM が比較 3 システムよりも再現率・F 値のほとんどにおいて上回り，総合的に見ると TVM は,「タップしながら歌唱する」という負荷の高さにもかかわらず，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音数・音高判定が実現可能であることが分かる．

ただし，被験者 A, E, F は，1 音ごとに正しくタップされなかったため結合音が多い．そして，A, F は結合音に起因する誤り音も多い．TVM では，結合音の音高は，結合音区間に含まれる音のうち，最も頻度の高い音高が採用される．また同一音高の連続箇所に限らずタップ区切りをしなければ結合音が発生するため誤り音と結合音が同時に発生しやすくなる．よって再現率あるいは適合率の精度低下が見られた．

しかし F 値で評価したところ，各被験者とも TVM が高いかあるいは同等となったため，TVM はより良好な性能を達成しているといえる．

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共に存在した．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かか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いえる．

5.4 楽器経験の有無のタップへの影響

提案手法（TVM）に必要なタップの能力が，楽器経験に影響されるかを評価した．まず楽器未経験者 A〜D 及び経験者 F〜I の 2 群に分けて，課題曲の TVM の結果比較を行う．被験者 E は楽器経験はあるがごく短く，どちらの群が妥当か判断しにくいので除いた．

テンポ自由歌唱では，楽器未経験者は再現率 98.7%，適合率 98.7%，経験者は同 99.7%，99.7%であった．これについて楽器未経験者と経験者の再現率及び適合率について t 検定を行ったところ，どちらも有意な差は見られなかった．また，再現率・適合率共に 100%の被験者が 5 名いたが，未経験者も含まれており，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

表 7 自由曲の変換結果 [歌唱条件：テンポ自由，歌詞歌唱，タップあり]
Table 7 Results of self-selected songs: [sung with own tempo, lyrics and taps].

被験者	全歌唱音(音)	正解(音)				上:誤り(音)/下:結合(音)				上:欠落(音)/下:結合(音)				余分(音)				再現率(%)				適合率(%)				F値			
		TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2	TVM	CMP	RYN	BP2
A*	120	**85**	82	**97**	92	15	12	14	5	20	26	9	23	0	10	15	9	70.8	68.3	80.8	76.7	**85.0**	**78.8**	**77.0**	86.8	77.3	73.2	78.9	81.4
						13	9	5	3	16	18	8	7																
B*	63	58	50	46	44	5	2	3	1	0	11	14	18	0	3	4	2	92.1	79.4	73.0	69.8	**92.1**	**90.9**	**86.8**	93.6	92.1	84.7	79.3	80.0
						0	1	0	0	0	7	3	7																
C*	61	52	43	37	14	9	4	3	3	0	14	21	44	0	9	0	0	85.2	70.5	60.7	23.0	**85.2**	**76.8**	**92.5**	**82.4**	85.2	73.5	73.3	35.9
						0	1	1	3	0	9	7	14																
D*	122	120	83	92	99	2	4	10	0	0	35	20	23	0	14	16	20	98.4	68.0	75.4	81.1	98.4	**82.2**	**78.0**	**83.2**	98.4	74.4	76.7	82.2
						0	2	1	0	0	30	19	16																
E	98	80	65	79	65	10	10	7	0	8	23	12	33	**8**	27	20	4	81.6	66.3	80.6	66.3	**81.6**	**63.7**	**74.5**	94.2	81.6	65.0	77.5	77.8
						1	1	2	0	7	9	12	11																
F	172	**155**	125	134	134	8	9	7	1	9	38	31	37	2	60	39	31	90.1	72.7	77.9	77.9	93.9	**64.4**	**74.4**	**80.7**	92.0	68.3	76.1	79.3
						5	3	0	0	9	26	27	21																
G	90	90	71	78	66	0	5	4	1	0	14	8	23	0	27	13	12	100	78.9	86.7	73.3	100	**68.9**	**82.1**	**83.5**	100	73.6	84.3	78.1
						0	2	0	0	0	8	8	14																
H	198	193	139	149	140	3	7	5	1	2	52	44	57	0	5	3	0	97.5	70.2	75.3	70.7	**98.5**	**92.1**	**94.9**	99.3	98.0	79.7	83.9	82.6
						2	6	1	1	2	36	42	43																
I	209	197	176	179	166	12	5	11	2	0	28	19	41	1	22	19	7	94.3	84.2	85.6	79.4	**93.8**	**86.7**	**85.6**	94.9	94.0	85.4	85.6	86.5
						0	3	1	1	0	17	16	17																
合計	1133	1030	834	891	820	64	58	64	14	39	241	178	299	11	177	129	85	90.9	73.6	78.6	72.4	93.2	**78.0**	**82.2**	**89.2**	92.0	75.7	80.4	79.9
						21	28	11	8	34	160	142	150																

注 1. “*” 付きの被験者は「楽器経験なし」と回答した被験者
注 2. 欠落音の下段は欠落音中の結合音数，誤り音の下段は誤り音中の結合音に起因する誤り音数を示す．また，誤り音と結合音由来の誤り音の差分は F0 推定ミス由来の誤り音数を示す．
注 3. 黒地白文字：4 システム中最もよい値を示す．

表 8　タップの有無による赤とんぼの被験者全体の再現率・適合率・F 値の比較（テンポ自由）
Table 8　Differences of the addition of tapping in the total values of recall, precision and F-value of “Aka tombo” (sung with own tempo).

	CMP		RYN		BP2	
	タップ有	タップ無	有	無	有	無
再現率	85.6	85.4	87.2	88.9	92.7	94.7
適合率	84.0	83.3	79.5	75.4	88.7	86.4
F 値	84.8	84.3	83.2	81.6	90.6	90.4

単位：%

表 9　タップの有無による赤とんぼの被験者全体の再現率・適合率・F 値の比較（BPM=120）
Table 9　Differences of the addition of tapping in the total values of recall, precision and F-value of “Aka tombo” (sung with BPM=120).

	CMP		RYN		BP2	
	タップ有	タップ無	有	無	有	無
再現率	83.8	86.1	87.5	81.7	78.5	79.0
適合率	86.9	86.7	84.7	77.1	92.1	92.1
F 値	85.3	86.4	86.1	79.3	84.8	85.0

単位：%

響を及ぼしにくいと見られる．

BPM=120 の歌唱では，未経験者は再現率 97.8%，適合率 96.6%，経験者は同 98.1%，98.1%であった．これについても楽器未経験者と経験者の再現率及び適合率について t 検定を行ったところ，どちらも有意な差は見られなかった．また，再現率・適合率共に 100%の被験者が 2 名いたが，1 名が未経験者であった．これらより多少速いテンポの入力であっても楽器経験の有無は影響を及ぼしにくいと考えられる．

次に課題曲の TVM の結果について表 1 の予備調査の結果も交えて評価した．

まず，表 1 の全 4 項目（音高聴取の結果は合計して使用）と全被験者のテンポ自由歌唱の正解音数とを重回帰分析した．楽器経験については，楽器経験があれば通常，リズムの知識や練習経験があると考えられるため，楽器に関係なく年数をそのまま用いることとした．複数の楽器経験がある場合は長い方を，範囲による回答の場合は長い方，1 年未満のものは月数を 12 か月で割った値を用いた．その結果，求められた重回帰式は，有意性が認められないものであった．

同様に BPM=120 の歌唱についても，重回帰式には有意性が認められなかった．これらの結果から，楽器経験とタップ能力の間には有意な相関が認められなかったことから，楽器経験はタップ能力に影響しないと思われる．

5.5　タップの有無の歌唱への影響

タップによって歌唱が不安定になるなどの影響があれば，判定精度にも何らかの影響が出る可能性がある．そこで，タップなしの歌唱による変換結果が得られる TVM 以外の 3 システムの課題曲の結果を用いて，タップあり（タップしながら歌唱したが，3 システムともタップ情報は処理に用いていない）とタップなしとで比較し，タップの歌唱への影響を調べた．

表 8 にテンポ自由歌唱の結果を示す．各システムについてタップの有無に分けて，全被験者の合計値を示す．全被験者の歌唱音数（母数）はタップありで 836 音，タップなしで 830 音であった．CMP, RYN, BP2 いずれも再現率，適合率共にタップの有無によらず同等の判定精度であった．よって，タップの有無はほとんど影響しないと考えられる．

表 9 に BPM=120 の歌唱の結果を示す．全被験者の歌唱音数（母数）はタップありで 837 音，タップなしで 835 音であった．BP2 は，タップの有無にかかわらず再現率・適合率共に大きな差は見られなかった．CMP では，自由テンポ時には同等だった再現率が，タップありの方がやや低くなった．RYN はタップありで再現率 87.5%，適合率 84.7%，タップなしで同 81.7%，77.1%であり，タップありが再現率・適合率共にタップなしを上回った．これは，被験者 E のタップなし歌唱時の誤り音が 35 音でタップあり歌唱時の 11 音に対して大きく増えているのが主因である．

以上から，総じて赤とんぼのような曲やテンポでは，タップの有無は歌唱にほとんど影響しないといえる．なお，BPM=120 の場合にタップの有無が若干影響する可能性が見られたが，必ずしもタップ有の場合に悪影響が出るわけではない．

5.6　全 体 考 察

TVM システムは，歌唱時の負荷や速いテンポなどでタップと歌唱のズレの発生はあるものの，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，欠落する音や不要な音の発生が抑制され，音数及び音高判定精度が向上することが示された．

楽器経験の有無のタップへの影響については，赤とんぼレベルの曲であれば，多少速いテンポの入力であっても大きく影響しないと見られることが分かった．また，タップの有無の歌唱への影響についても，赤とんぼレベルの曲の場合，入力テンポが速くなると多少影響が出る可能性があるものの，必ずしもタップが悪

影響を及ぼすわけではなく，総じてタップの有無の影響は小さいことが分かった．

我々は，これまでに文献 [22] において市販の「タタタ歌唱」システムに自由歌唱を入力して比較実験を行っている．文献 [22] で用いたのは，今回提案した手法と比べて無発声検知機構や精度の高い F0 推定法を採用していない，性能が劣る手法であったが，「タタタ歌唱」を必要とするシステムに対する優位性を示している（提案手法は再現率 65.9%，適合率 70.2%，比較システムは同 24.0%，36.8%）．この結果と合わせて，総じて TVM は十分な有用性があると考えられる．

6. むすび

本論文では，Voice-to-MIDI システムの音数・音高の判定精度向上のために，メロディリズムの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な音数・音高判定手法を提案した．次に提案手法について，歌詞歌唱などの任意発音の歌唱を許容する既存 VtoM システムとの変換精度の比較，楽器経験の有無のタップへの影響やタップの有無の歌唱への影響の評価を行った．その結果，タップの付加により音数抽出の正確さが増し，それが音高判定の精度向上にも寄与したことを示した．

今後，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することや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．また，音楽療法支援への適用について実践的応用例として進めていく予定である．

なお現段階では，実装の容易性のために出力には MIDI Note No. を利用している．このため，システムの名称も Voice-to-MIDI としている．しかしながら，これが唯一の実装形態というわけではなく，将来的には，より広く音声を音楽的な表現に変換するシステム (Voice-to-MusicalExpression) を実現することを目指している．

謝辞 文献 [10] について，プログラムの提供及び比較評価への使用を快諾頂いた，Matti Ryynanen 氏及び Anssi Klapuri 博士に感謝の意を表します．

また，多忙な中，評価実験に参加頂いた，被験者の皆様に感謝の意を表します．

文　献

[1] YAMAHA，XGworks ST，浜松，2003.
[2] INTERNET，SingerSongWriter Lite6.0, 大阪，2008.
[3] MakeMusic Inc., Finale2010, USA, 2009.
[4] 野口義修，“詞先・メロ先作曲術，” 作曲本，pp.109–118，シンコーミュージック・エンタテインメント，東京，2005.
[5] 奥平ともあき，“詞先・曲先，” 誰にでもできる作曲講座，pp.20–21，ドレミ楽譜出版社，東京，2003.
[6] C. Oshima, N. Itou, K. Nishimoto, N. Hosoi, K. Yasuda, and K. Nakayama, “An accompaniment system for healing emotions of patients with dementia who repeat stereotypical utterances,” Proc. 9th Int’l. Conf. Smart Homes and Health Telematics, 2011.
[7] 新原高水，今井正和，井口征士，“歌唱の自動採譜，” 計測自動制御学会論文集，vol.20, no.10, pp.68–73, 1984.
[8] C.C. Toh, B. Zhang, and Y. Wang, “Multiple-feature fusion based onset detection for solo singing voice,” Proc. ISMIR 2008, 2008.
[9] M. Ryynanen and A. Klapuri, “Modelling of note events for singing transcription,” Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio, 2004.
[10] M. Ryynanen and A. Klapuri, “Automatic transcription of melody, bass line, and chords in polyphonic music,” Computer Music Journal, vol.32, no.3, pp.73–86, 2008.
[11] T. Kageyama, K. Mochizuki, and Y. Takashima, “Melody retrieval with humming,” Proc. ICMC 1993, pp.349–351, 1993.
[12] A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith, “Query by humming: Musical information retrieval in an audio database,” Proc. ACM Multimedia’95, San Francisco, California, Nov. 1995.
[13] L. Prechelt and R. Typke, “An interface for melody input,” ACM Trans. Computer-Human Interaction (TOCHI), vol.8, no.2, pp.133–149, 2001.
[14] N. Kosugi, Y. Nishihara, T. Sakata, M. Yamamuro, and K. Kushima, “A practical query-by-humming system for a large music database,” Proc. 8th ACM Intl. Conf. Multimedia, pp.333–342, Marina del Rey, California, 2000.
[15] 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，” 情処学音声言語情報処理研報，SLP-47, pp.71–76, 2003.
[16] 岩田憲治，渡邉康司，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，”2006 音響秋季講論集，1-2-23, 2006.
[17] Wildcat Canyon Software Inc., Autoscore 2.0, 1999.
[18] 半田伊吹，木下智義，武藤　誠，坂井修一，田中英彦，“マン・マシン協調による採譜システム，” 情処学音楽情報科学研報，MUS-34, pp.21–26, 1999.
[19] 野ばら社編集部，童謡，p.68, 野ばら社，東京，1994.
[20] 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，pp.718–723, 1983.
[21] 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，” 情処学エンタテインメントコンピューティング研報，2006-EC-5, vol.2006, pp.43–48, 2006.
[22] N. Itou and K. Nishimoto, “A voice-to-MIDI system

for singing melodies with lyrics," Proc. Intl. Conf. ACE'07, pp.183–189, Salzburg, Austria, 2007.
[23] 金澤正剛（監修），"記号表," 新編音楽小辞典，p.439, 音楽之友社，東京，2004.
[24] 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," 音響学音楽音響研資，vol.23, no.5, pp.95–100, 2004.
[25] 河合楽器製作所，Band Producer 2, 浜松，2008.

（平成 24 年 7 月 14 日受付，10 月 25 日再受付）

伊藤　直樹　（正員）

2011 北陸先端科学技術大学院大学知識科学研究科博士後期課程単位取得満期退学．同年インターメディアプランニング（株）入社．音楽情報処理を中心としたエンタテインメントシステムのほか，モチベーション支援，意思共有支援に興味をもつ．ICOST2011 Best Multi-Disciplinary Paper Award，GLOBAL HEALTH 2012 Best Paper Award 受賞．情報処理学会会員．

西本　一志

1987 京都大学大学院工学研究科機械工学専攻博士前期課程了．同年松下電器産業（株）入社．1992（株）ATR 通信システム研究所出向．1995（株）ATR 知能映像通信研究所客員研究員．1999 より北陸先端科学技術大学院大学助教授．2007 より教授．2000～2003 科学技術振興事業団さきがけ研究 21「情報と知」領域研究員兼任．1999 年度情報処理学会坂井記念特別賞，1999 年度人工知能学会論文賞，ACM Multimedia 2004 Best Paper Award，ICOST2011 Best Multi-Disciplinary Paper Award，GLOBAL HEALTH 2012 Best Paper Award 等受賞．IEEE computer society，ACM，情報処理学会，人工知能学会，ヒューマンインタフェース学会各会員．博士（工学）．