Vol.2014-HCI-156 No.13
2014/1/16

スマートデバイスにおいて楽曲の音響特徴を利用し楽曲を高速選択する手法の提案と評価

萩原 崇文[†1,a)]　赤池 英夫[†1]　角田 博保[†1]

概要：近年，Walkman や iPod Touch 等のような，汎用 OS を搭載した音楽プレイヤー・スマートフォンで音楽を聴く機会も増加してきた．これらの端末は数千曲もの楽曲を保存し再生することができる．しかし，一方で，端末内の楽曲の数が増えてくると，曲やアルバム等の選曲画面から聴きたい曲を選ぶのに時間がかかってしまうという問題が発生する．
そこで，本研究では，携帯端末用の直感的な選曲システムとして，端末の画面をタップする時のリズム・強さ，およびタップ位置による音程指定の情報を利用して聴きたい曲を絞り込み，より素早く選曲するシステムを設計・評価する．
端末で，音高変化と感情の強さを入力し，予め作成された楽曲側のデータと比較を行う．入力データにより近い楽曲を選曲できるようにする．
本稿では，主に，楽曲のボーカルの音響信号から音高変化を抽出し楽曲データを作成する方法，入力されたリズムや音高変化と比較し，候補を絞り込む方法やその結果について報告する．

1. 研究背景

近年，Walkman や iPod Touch 等の音楽プレイヤーが普及し，Android や iOS のような汎用 OS を搭載している場合も多い．また，これらの OS を搭載したスマートフォンで音楽を聴く場合も増えてきた．これらの端末は 10G バイト以上のデータを保存することが多いため，端末 1 台に数千曲もの楽曲を保存し再生することが可能になっている．一方で，楽曲が増えてくることにより，曲やアルバム等の選択画面から聴きたい曲を選ぶ際にスクロールの回数が増え，操作の手間が増えてしまう問題が発生する．また，聴きたい曲やアルバムの名前をはっきりと覚えておらず，その曲を探すのに時間がかかってしまうという問題が発生する．また，楽曲を素早く検索する手法として鼻歌による楽曲検索が挙げられるが，電車内などの公共空間での使用は難しい．

そこで，本研究では携帯端末用の直感的な選曲システムとして，端末の画面を操作する時のタップしたボタンやそのリズム・強さの情報を利用して聴きたい曲を絞り込むことにより，聴きたい楽曲をより素早く選曲するシステムを設計・評価する．また，複数の条件を組み合わせることにより，選曲する曲の候補を絞り込みやすくする．

[†1] 現在，電気通信大学大学院 情報理工学研究科 情報・通信工学専攻
Presently with Department of Communication Engineering and Informatics, Graduate School of Informatics and Engineering, The University of Electro-Coummunications
[a)] tahagi@gulf.cs.uec.ac.jp

2. 関連研究

池谷ら [1] の研究では，1 個のボタンをリズミカルにタップする事により，音楽を検索するアルゴリズムを提案している．MIDI から検索用のデータとして作成した 2500 曲中，意図した曲が 5 位以内に表示される確率が 75%という結果になっている．

Bandera ら [2] の研究では鼻歌により楽曲検索を行うことができる．検索用のデータを楽曲波形から生成することにより，140 曲からポップスやロックの曲を探す場合において，意図した楽曲が 5 位以内に表示される確率が 45.10%であると示されている．

タップにリズム以外の情報を付加したものとして，石山ら [3] の研究が挙げられる．音声コマンドをリズムに置き換え，音声コマンドを発声するときのアクセントの強弱を付加することにより，リズミカルに画面をタップすることにより携帯端末を操作できるようにした．アクセントの強弱の識別は端末のサイドボタンを押すことで行なっており，直感的ではない．

Rui ら [4] の研究では，様々な楽曲の音色等の音響特徴からシグネチャを作成し，音響特徴が近い楽曲同士で自動的にプレイリストを作成するシステムについて示されてい

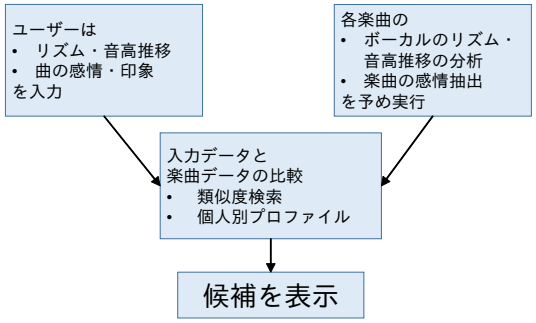

図 1 システムの概要図
Fig. 1 Overview of system.

る．しかし，単に音色などの音響特徴を用いた照合方法では，特定の曲を探す目的でユーザーがタップ操作によって検索クエリを入力するには不向きであると考えられる．

3. 提案手法

本研究では，スマートフォンを使用してタップ入力により選曲を行うシステムを提案する．端末の画面にボタンが表示される．ユーザーは，聴きたい曲のリズムに合わせて端末の画面に表示されたボタンをタップする．その時，リズム・曲の感情や印象・メロディの音高推移を利用する．メロディの音高推移とリズムはボタンをタップする順番やタイミングによって表現する．音高推移とは，楽曲のある音が，前の音よりも低くなったか，変わらないか，あるいは高くなったかといった変化のことを言う．また，曲の感情や印象は，タップの強弱によって表現する．例えば，激しい曲やポジティブな曲では強くタップし，静かな曲やネガティブな曲は弱くタップする．これを感情レベルと定義する．

4. 設計方針

システムの概要図を**図 1** に示す．

選曲対象の楽曲に対して，波形データから予めボーカルの発音タイミング・音高推移の分析および楽曲の感情レベルの決定を行ない，選曲用データベース (DB) を自動的に生成しておく．ユーザーはタップ操作で聴きたい曲の情報を入力する．入力されたデータは選曲用 DB と照合し，類似度を比較することにより楽曲を絞込み，候補を表示する．また，ユーザー別に選択する曲の傾向を学習し，聴きたい曲がより候補に現れるようなシステムを実装する．

4.1 入力画面

ユーザーは，入力画面でリズムに合わせて音高推移に合わせたボタン入力を行う．画面上に，**図 2** のように左右に数個のボタンを配置する．ユーザーは音高推移に合わせてボタンの入力を行う．

前の音より音高が上がる時は，1 つ前に押したボタンの右隣をタップする，但し，1 つ前が右端のボタンだった場合は同じ右端のボタンを押す．反対に，前の音より音高が下がる時は，1 つ前に押したボタンの左隣をタップする．但し，1 つ前が左端のボタンだった場合は，同じ左端のボタンを押す．1 つ前の音と同じ音の場合は，同じボタンをタップする．また，入力時に，感情レベルの強さに応じてボタンをタップするときの強弱をつける．これらの操作によって，ボーカルの発音リズムと音高推移・楽曲の感情レベルを同時に入力する．図 2 では，ボタンが 4 つであるが，最適なボタンの個数も検討する．

入力後，システムは入力された情報に類似している曲の候補を提示する．

図 2 選曲時の入力画面 (ボタン 4 つの場合)
Fig. 2 The display layout for music selection. (4 buttons)

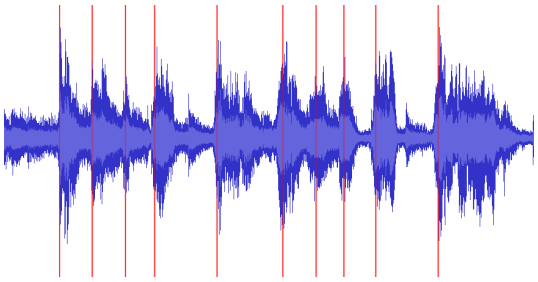

図 3 ボーカル分離後の音声波形の一部と発音時刻
Fig. 3 Vocal separated waveform on a music and Onset time

4.2 リズム・音高推移の分析と入力

音高推移については，楽曲の波形からボーカルを抽出し，その音高の変化を利用する．ボーカルの抽出には，REPET-SIM[6], [7] を用いる．REPET-SIM は，音響信号から楽器のベース音などの繰り返される音響成分を除去し，ボーカルなどの成分を高精度で分離することができる．

ベース音等を除去し，ボーカルが残った音響信号の波形を**図 3** とした時，声を出した時に図中の縦線の部分のように波形が急峻になる．この波形の急峻な部分を認識し，発音時刻を取得していく．

K 回発音があったとすると，k 回目の発音時刻を s_{mk} とすると，それらの列を

$$\mathbf{S}_m = (s_{m1}, \cdots, s_{mk}, \cdots, s_{mK}) \qquad (1)$$

と表すことができる．

一方，曲 m のボーカルを抽出した音響信号は，音響信号解析ライブラリ MARSYAS[8] の AubioYin モジュールにより，ボーカルのピッチを検出する事ができる．AubioYin モジュールは，Yin のアルゴリズム [9] によって声のピッチを検出することができる．

各発音時刻 s_{mk} においてピッチを検出し，その値を p_{mk} の列

$$\mathbf{P}_m = (p_{m1}, \cdots, p_{mk}, \cdots, p_{mK}) \tag{2}$$

として得られる．

聴きたい曲のリズム・音高推移を端末で入力する時は，ユーザーがその曲の解析されたフレーズと同じ部分を入力する事によって行う．

4.3 感情レベルの分析と入力

三好ら [5] の研究で，楽曲に適切な印象値を自動的に付与する技術の評価が行われている．音量や音色・リズムや和音等の特徴からニューラルネットワークを用いて，楽曲印象値を 7 段階で自動的に付与することが可能である．

本研究では，印象値の情報を利用して，各楽曲のネガティブ・ポジティブの度合いを決定し，感情レベルとすることを検討する．特に，タップの強さにより感情レベルを入力する場合，ユーザーによってより分かりやすい感情レベルの指標が必要になる．そのため，タップによる強弱がユーザーにとって直感的により分かりやすいような感情レベルの定義を検討する．例えば，曲 m では，k 回目の発音時刻における感情レベルを f_{mk} として，

$$\mathbf{F}_m = (f_{m1}, \cdots, f_{mk}, \cdots, f_{mK}) \tag{3}$$

のように曲ごとの感情レベルを決定する．

タップの強弱は，端末による指の接触面積の検出，またはマイクや加速度センサで検出する．

4.4 入力データと楽曲データの比較

入力されたデータと選曲用 DB のデータを基に，類似度を比較する．l 回目のタップの時刻を ${s'}_l$ とした時，L 回タップした時の各タップ時刻の列を，

$$\mathbf{S}' = ({s'}_1, \cdots, {s'}_l, \cdots, {s'}_L) \tag{4}$$

と表す．l 回目のタップにおける音高推移を ${p'}_l$ とした時，音高推移の列を

$$\mathbf{P}' = ({p'}_1, \cdots, {p'}_l, \cdots, {p'}_L) \tag{5}$$

と表す．但し，各 ${p'}_l$ は，音高が高くなったら $+1$，変わらない場合は 0，音高が低くなったら -1 を表す．l 回目のタップにおける感情レベルを ${f'}_l$ とした時，感情レベルの列を

$$\mathbf{F}' = ({f'}_1, \cdots, {f'}_l, \cdots, {f'}_L) \tag{6}$$

と表す．

ここで，楽曲 m の発音時刻 $\mathbf{S_m}$ について，各発音時刻同士の間隔をできるだけ小さい整数比になるようにし，それを拍数とする．拍数を t_{ml} とすると，

$$\mathbf{T}_m = (t_{m1}, \cdots, t_{ml}, \cdots, t_{mK-1}) \tag{7}$$

とおくことができる．

また，入力データにおけるタップ時刻 $\mathbf{S}'$ についても同様に，各発音時刻同士の間隔をできるだけ小さい整数比になるようにし，それを拍数 ${t'}_l$ とすると，

$$\mathbf{T}' = ({t'}_1, \cdots, {t'}_l, \cdots, {t'}_{L-1}) \tag{8}$$

とおくことができる．

入力されたデータは，選曲用 DB のデータと比較を行う．

入力データの $\mathbf{T}', \mathbf{P}', \mathbf{F}'$ と曲 m の $\mathbf{T}_m, \mathbf{P}_m, \mathbf{F}_m$ について，$\mathbf{T}'$ と $\mathbf{T}_m$ で拍がより正確に一致し，$\mathbf{P}'$ と $\mathbf{P}_m$ で音高の変化がより一致する部分，$\mathbf{F}'$ と $\mathbf{F}_m$ で感情レベルがより一致する部分において，その類似度をその楽曲 m のスコアとする．

各楽曲のスコアを比較し，よりスコアが高い楽曲が，より入力された楽曲のタップ操作に近い楽曲となる．

4.5 個人別プロファイル

ユーザーのタップのタイミングや楽曲からの選曲用データの取得精度等によって楽曲選択の精度に差が出る可能性が発生する．そのため，ユーザー別によく選曲する楽曲の傾向や Last.fm [10] から取得できる類似・関連アーティスト等の情報も利用して，その結果を選曲候補に反映させることを検討する．

5. 予備実験

5.1 目的

端末画面のボタンによる入力を検討するにあたり，どのようなインタラクションにより音高推移を入力すると，ユーザーが曲の音高の推移を認知し，よりタップ入力が行いやすいのかを検証するために予備実験を行った．

5.2 タスク

この実験では，2〜4 小節程度の単音の曲を流し，その音高推移を音に合わせてスマートフォン画面に入力できるか実験を行った．

各曲ごとに，被験者は最初に 1 回だけどのような曲なのかを聞いた．その後，最初に流れた曲と同じ曲が流れるので，それに合わせて音高が下がったか，変わらないか，あるいは上がったかを端末の画面のボタン (5.3 節を参照) を押すことにより入力した．1 曲につき，入力は 3 回繰り返

した．

各被験者は練習を 1 曲行い，その後 10 曲で実験を行った．この手順を 1 セッションとし，1 人 2 セッション行った．

5.3 画面のタップ操作

予備実験では，3 種類のタップ操作を使用した．被験者別にタップ操作のタイプを振り分けた．

タイプ 1(図 **4**) は，ボタン 3 つで構成されており，音高が下がったら左のボタンを，変わらなかったら中央のボタンを，音高が上がった場合は右のボタンを押す．

タイプ 2(図 **5**) は，ボタン 2 つで構成されており，音高が下がったら左のボタンを，音高が上がった場合は右のボタンを押す．音高が変わらなかったらどちらのボタンを押しても良い．

タイプ 3(図 **6**) は，ボタン 2 つで構成されており，音高が下がった，または上がったら左のボタンを，音高が変わらなかった場合は右のボタンを押す．

5.4 実験環境

本実験用のスマートフォンに，Sony XPERIA SO-02D または Sony XPERIA SO-03D を使用した．

被験者には，椅子に座り，スマートフォンを横画面で片手で把持するように指示した．また，画面を操作する側の指は複数同時に使用してもよいとした．被験者は 1 セッションで練習 1 曲，本番 10 曲の操作を行った．30 分以上間を空けてから 2 セッション目を行うように指示した．被験者ごとにタップ操作のタイプを指定し，2 セッション共に同じタイプで実験を行わせた．

使用した曲は，童謡や唱歌 11 曲 (練習用の 1 曲を含む) である．

被験者は，本大学の学生 9 名 (女性 1 名) である．タップ操作のタイプ 1 種類につき 3 名ずつを割り当てて実験を行った．

5.5 結果

被験者別に，各セッションで音高推移を正しく入力できた割合 (正答率) を**表 1** に示す．タイプ 2 の正答率が最も高く，また，タイプ 2 とタイプ 3 では，2 セッション目の正答率が 1 セッション目よりも良くなっていることがわかった．

5.6 考察

以上の実験より，タイプ 2 のタップ操作が音高推移を入力しやすいと考えられる．これは，音高が下がったら左側を押して，音高が上がったら右側を押せばよいという単純な配置であったため，鍵盤楽器を弾くような感覚で直感的に入力しやすいためと考えられる．

この予備実験では，タイプ 2 は，ボタンが 2 つの場合で実験を行ったが，そのボタンの数を任意の個数 n にした場合について今後比較検討する．

表 1　予備実験の結果 (被験者別)

Table 1　Result of a preliminary test (By subjects).

タップ操作のタイプ	被験者	正答率 (%)	
		セッション 1	セッション 2
タイプ 1	A	47.0	46.7
	B	19.1	18.2
	C	27.9	25.0
	A〜C 全体	31.3	30.0
タイプ 2	D	85.1	88.7
	E	84.6	92.7
	F	73.1	84.6
	D〜F 全体	80.9	88.7
タイプ 3	G	65.0	75.0
	H	83.4	89.8
	I	71.7	73.3
	G〜I 全体	73.4	79.4

6. おわりに

本研究では，端末の画面をタップする時のリズム・強さ・音程指定によって楽曲を選曲するシステムを提案した．そして，端末でどのようにボタンの機能を割り当てると入力が行いやすいかを評価した．

今後，端末の入力画面や感情レベルの決定方法，および入力されたデータと楽曲側のデータをより正確に比較する方法を検討する．最終的に，これらの方法で正しく選曲できるか，また，ユーザーにとって選曲操作が行いやすいかを評価する．

本研究は，RWC 研究用音楽データベース (ポピュラー音楽) を利用した．

参考文献

[1] 池谷 直紀, 服部 正典, 梅木 秀雄, 大須賀 昭彦: リズム入力インタフェース「タタタップ」による大規模音楽検索, 情報処理学会研究報告,2005-HI-113,**Vol.2005**,pp.27-33 (2005).

[2] Bandera,C., Barbancho,A., Tardón,L.J, Sammartino,S., Barbancho,I.: HUMMING METHOD FOR CONTENT-BASED MUSIC INFORMATION RETRIEVAL, ISMIR, pp.49-54 (2011).

[3] 石山 英貴, 高橋 伸, 田中 二郎: コマンドリズムを用いたタップ入力による携帯端末操作手法, 情報処理学会 インタラクション 2013,2013-Interaction (1EXB-35),pp.270-277 (2013).

[4] Cai,R., Zhang,C., Zhang,L., Ma,W.: Scalable Music Recommendation by Search, Proceedings of the 15th international conference on Multimedia,pp.1065-1074 (2007).

[5] 三好 真人, 柘植 覚, ChogeKipsang,H., 尾山 匡浩, 伊藤 桃代, 福見 稔: 音楽検索のための楽曲印象値の自動付与手法, 情報処理学会研究報告,2011-MUS-89,**Vol.2011**,pp.1-6 (2011).

[6] Rafii,Z., Pardo,B.: MUSIC/VOICE SEPARATION US-

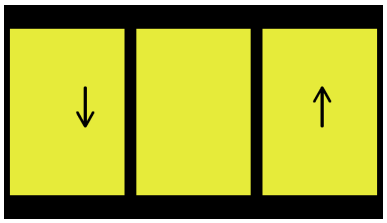

図 4 操作画面:タイプ 1
Fig. 4 Screen layout of the type 1 operation

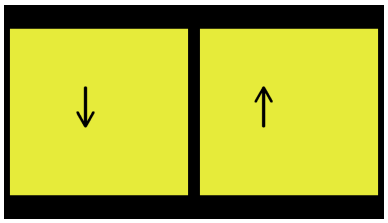

図 5 操作画面:タイプ 2
Fig. 5 Screen layout of the type 2 operation

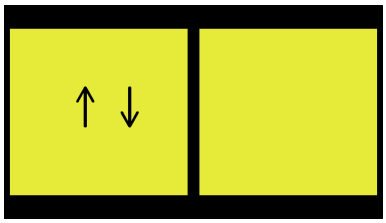

図 6 操作画面:タイプ 3
Fig. 6 Screen layout of the type 3 operation

ING THE SIMILARITY MATRIX, ISMIR, pp.583-588 (2012).
[7] Rafii,Z., Pardo,B.: Online REPET-SIM for real-time speech enhancement, ICASSP, pp.848-852 (2013).
[8] Marsyas, `http://marsyas.info/`
[9] Cheveigné,A., Kawahara,H.: YIN, a fundamental frequency estimator for speech and musica, J. Acoust. Soc. Am. 111,1917 (2002)
[10] Last.fm, `http://www.lastfm.jp`