2007－MUS－72 (9)
2007／10／12

Webから収集した楽曲を説明するテキストと楽曲の音響特徴量との関連づけに関する検討

高橋 量衛† 大石 康智† 武田 一哉†
†名古屋大学大学院情報科学研究科
†{takahasi, ohishi}[at]sp.m.is.nagoya-u.ac.jp
kazuya.takeda[at]nagoya-u.jp

あらまし 本研究では，ユーザが閲覧しているWebページにふさわしいBGMを，自動的に選曲するという新しい楽曲検索（推薦）システムを提案する．閲覧ページのテキストに含まれる語彙の共起から，それらの語彙に関連する楽曲の検索，推薦を行う．そのためには，語彙の共起に基づく特徴空間と楽曲の音響的特徴空間とを関連付ける必要がある．我々は，線形変換を用いてこの関連付けを実装した．さらに，Webから収集した楽曲のレビューのようなテキストデータと，その楽曲の音響特徴量を用いて，関連付けの性能評価実験を行った．その結果，各楽曲に対し1つのレビューを利用するより，曲名とアーティスト名を含むWebページを複数利用した場合に関連付け性能が高いことを確認した．

Association between song reviews collected from Web and acoustic features

Ryoei TAKAHASHI† Yasunori OHISHI† Kazuya TAKEDA†
†Graduate School of Information Science, Nagoya University

Abstract A new music information retrieval application, WEB-BGM that automatically selects and plays the background music for the web page under browsing is proposed. In order to find for the song that is 'near' to the browsing page, the song is needed to be located in the document space. However, in general, the documents relevant to the song, e.g. reviews of the song, are not available for each songs. Therefore, we train a matrix that transforms a document vector onto acoustic space so that to find 'nearest' song to the web page in the acoustic space. The feasibility of the idea is confirmed through preliminary experiments using song reviews and Web pages including the song title and artist name.

1 はじめに

インターネットを利用して大規模楽曲データベースにアクセスし，ユーザが大量の楽曲を所有できるようになった．大量の楽曲を管理し，ユーザが効率よく検索して鑑賞するための技術が必要とされている．一般的な楽曲の試聴方法といえば，ユーザが曲名，アーティスト名を選択するもの，アルバムを単位として視聴するもの，あらかじめ好みの楽曲からなるプレイリストを手作業，もしくは自動的に作成して試聴するものがある．

また，「明るい」や「静かな」のような特定の感性語や音楽的な用語を，あらかじめ楽曲と関連付けておくことによって，それらの語彙から楽曲を検索する方法も提案されている．熊本ら [1] は，聴取実験に基づいて，楽曲の曲調を感性語で表現し，重回帰分析によって楽曲の音響的特徴との関連付けを行った．またTurnbullら [2] は，楽曲のレビューから音楽に関係する語彙を手作業で選定し，楽曲の音響的特徴との関連付けを確率的な手法を用いて行った．その結果，音楽的な語彙に基づく楽曲検索，また，楽曲に対して語彙を自動アノテーションすることが可能となった．

Whitman ら [3] はレビューに出現する語彙と音響的特徴との関連付けを識別器を用いて行い，楽曲とは無関係な語彙を取り除くことを試みている．以上のように，特定の語彙と楽曲の音響的特徴との関連付けについては従来行われていた．一方，Slaney[4] は楽曲ではなく実環境における音と，それを説明した文章との関連付けを試みている．

本研究では，ユーザが Web ページを閲覧しているときに，自動的に BGM を流す新しい楽曲検索（推薦）システムを提案する．つまり，閲覧している Web ページのテキストに含まれる複数の語彙の共起関係を利用して，その閲覧ページにふさわしい楽曲を自動選曲することを考える．図 1 に示すように，楽曲の書誌情報や歌詞，楽曲を解説したレビューなどの，テキストに出現する語彙の共起によって，楽曲と語彙との関係を特徴づける空間が構築される．この特徴空間を利用して，様々な語彙に基づいて楽曲を検索することは可能である．しかし，この特徴空間だけでは，楽曲の音響的特徴に基づく類似性が必ずしも反映されていない．なぜなら，音響的に類似した楽曲であっても，そのレビューでは，様々な語彙を活用して楽曲が解説されるためである．そこで，さらに楽曲の音響的特徴の類似性が表現された空間を用意し，この空間と語彙の共起関係を特徴づけた空間とを関連付ける手法を検討する．この関連付け手法によって，閲覧している Web ページの語彙に基づいて楽曲が検索され，さらに音響的に類似した楽曲をも，選曲することができるようになる．また，レビューが存在しない楽曲であっても，関連付け手法によって音響的特徴の類似性から選曲候補として挙げることも可能である．さらに図 1 を逆方向にたどることによって，楽曲を入力すれば，音響的特徴空間と語彙の特徴空間との関連付けから，その楽曲を解説するのに適した語彙集合を出力し，最終的に文章 (レビュー) を自動生成させるという応用も考えられる．

本稿では，この語彙の共起に基づく特徴空間と音響的特徴空間の表現方法，これら 2 つの特徴空間を関連付けるための手法について議論する．また，楽曲に関連する語彙を含むテキストデータとして，音楽ダウンロードサイトにおけるレビュー，検索エンジンを用いて収集したアーティスト名と曲名を含む Web ページを利用し，提案手法による関連付け性能について評価実験を行う．2,650 曲の試聴曲とレビューを利用し再

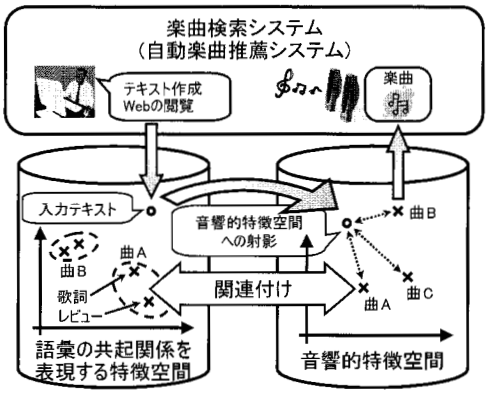

図 1　楽曲を解説するテキストと音響的特徴との関連付け手法を利用した楽曲検索（推薦）システム

現率，適合率による評価を行った．また，各楽曲に対し単一のレビューより曲名とアーティスト名を含む Web ページを複数利用すると関連付け性能が高いことを確認した．

以下，第 2 章では，語彙の共起に基づく特徴空間と音響的特徴空間との関連付け手法について述べる．第 3 章では，様々なテキストデータを利用して，その関連付け性能の評価実験を行う．第 4 章では，まとめと今後の課題について述べる．

2　楽曲を解説したテキストと音響的特徴との関連付け手法

楽曲を解説したテキストに出現する語彙の頻度分布に基づいて，テキストをベクトルで表現する（文書ベクトルと呼ぶ）．また，楽曲全体を信号処理することによって得られる音響的特徴量の頻度分布に基づいて，楽曲全体の音響的特徴をベクトルで表現する（音響ベクトルと呼ぶ）．最終的に，この 2 つの特徴ベクトルを線形変換によって関連付けるための手法を提案する．

2.1　TF-IDF を利用したテキストの特徴抽出

楽曲 j を解説したテキストを，文書ベクトル $\mathbf{x}_j$ で表現する．この文書ベクトル $\mathbf{x}_j$ の i 次元目の要素 $x_{i,j}$ は，形態素 t_i に関して以下の式で計算される TF-IDF (term frequency - inverse document frequency) に

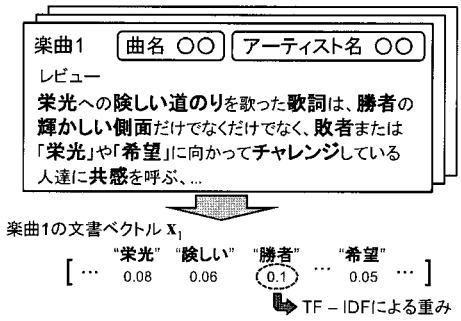

図2　文書ベクトルの作成

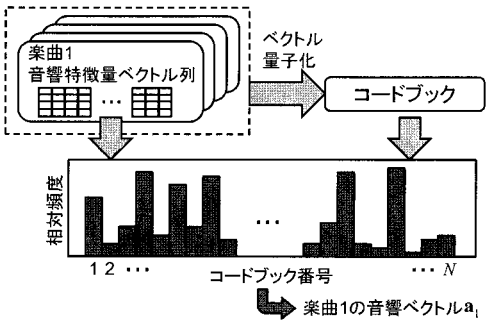

図3　音響ベクトルの作成

よる重みとする．

$$x_{i,j} = \frac{tf_{i,j}}{\sum_i tf_{i,j}} \times \log \frac{J}{df_i} \tag{1}$$

ここで，楽曲 j を解説したテキストにおける形態素 t_i の出現頻度を $tf_{i,j}$，すべてのテキストのうち，形態素 t_i を含むテキストの数を df_i，楽曲の総数を J とする．

さらに，文書ベクトルの集合を行列 $X = (\mathbf{x_1}, \cdots, \mathbf{x}_j \cdots, \mathbf{x}_J)$ と記述する．X は $I \times J$ の行列であり，I はテキストに出現する形態素の総数である．したがって，行列 X は 0 の要素を多く含むスパースで高次元の行列となる．そこで，高次元では別々の次元で扱われる語彙を，低次元で意味的に関連した語彙としてひとつの次元に縮退させるために，以下のように行列 X の特異値分解を行う [5]．

$$X = USV^T \tag{2}$$

ここで，S は $J \times J$ の非負要素の対角行列であり，対角要素は絶対値の降順に並んでいるものとする．直交行列 U のうち，絶対値の大きな特異値に対応する第 1 列から第 k 列を取り出した行列 U_k を用いて，I 次元の文書ベクトル $\mathbf{x}_j$ を以下のように k 次元に削減することができる．

$$\mathbf{t}_j = U_k^T \mathbf{x}_j \tag{3}$$

以後，この次元削減した $\mathbf{t}_j$ を文書ベクトルとして利用する．

2.2　スペクトル形状を利用した楽曲の音響的特徴抽出

楽曲の音響的特徴を表現するために，音響信号を短時間フーリエ変換 (STFT) して得られる短時間スペクトルから，以下の 5 つの特徴量を計算する．

1) スペクトル重心

以下のようにスペクトルの重心を計算する．

$$C_t = \frac{\sum_{k=1}^{K} M_t[k] * k}{\sum_{k=1}^{K} M_t[k]} \tag{4}$$

ここで，$M_t[k]$ は，t 番目のフレームをフーリエ変換して得られるスペクトルの，k 番目の周波数ビンにおける振幅値である．スペクトル重心は，スペクトルの形状を表現する尺度として利用されている [6]．重心の値が大きければ，より多くの高周波数成分を含むため，”明るい”音色をもつことに対応する．

2) スペクトルロールオフ

スペクトル分布の全帯域の 85% を占める周波数 R_t として以下のように定義される．

$$\sum_{k=1}^{R_t} M_t[k] = 0.85 * \sum_{k=1}^{K} M_t[k] \tag{5}$$

このロールオフもスペクトルの形状を表す一つの尺度とされる [6]．

3) スペクトルフラックス

スペクトル分布を正規化し，以下のように時間的に隣り合う振幅値の 2 乗誤差を計算する．

$$F_t = \sum_{k=1}^{K} (N_t[k] - N_{t-1}[k])^2 \tag{6}$$

ここで，$N_t[k]$ と $N_{t-1}[k]$ は t 番目のフレームと 1 つ前の $t-1$ 番目のフレームにおける正規化されたスペクトルの，k 番目の周波数ビンにおける振幅値である．スペクトルフラックスは，局所的なスペクトルの変化を表す尺度である [6]．

4) スペクトルフラットネス

t 番目のフレームのスペクトル分布の周波数帯域 B_i における幾何平均と算術平均との比を以下のように

計算する．

$$SFM_{t,i} = \frac{\left\{\prod_{k\in B_i} P(k)\right\}^{1/K_{B_i}}}{\frac{1}{K_{B_i}}\sum_{k\in B_i} P(k)} \quad (7)$$

ただし，$P(k)$ はパワースペクトル，k は周波数ビン，K_{B_i} は周波数帯域 B_i における周波数ビンの総数を表す．周波数帯域 B_i の帯域幅は次式のようにオクターブの 1/4 である．

$$1000 \times 2^{0.25\times(i-0.8)} < B_i < 1000 \times 2^{0.25\times(i+1-0.8)} \quad (8)$$

これより，250Hz〜16kHz の間で $i = 24$ 個の帯域となる．

5) ゼロ交差数

信号の時間波形がゼロと交差する回数を計算する．

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \quad (9)$$

ここで $sign$ 関数は，正であれば 1 を，負であれば 0 を返すものとする．$x[n]$ は，t 番目のフレームにおける時間波形を表す．ゼロ交差数は信号の雑音度を測る尺度である [6].

以上の特徴量の抽出には Marsyas[7] を利用した．さらにこれらの特徴量の動的変動成分として，前後 2 フレーム分の計 5 点の直線回帰によって得られる回帰係数を計算した．最終的に，以上の特徴量をすべてまとめて，t 番目のフレームにおける音響特徴量ベクトル $\mathbf{v}_t$ とする．

図 3 に示すように，全楽曲から求めたこの音響特徴量ベクトル $\mathbf{v}$ の集合を L 個のクラスにベクトル量子化するために，各セントロイドを表すコードブックを求める．次に各楽曲ごとに，音響特徴量ベクトルの集合をコードブックに基づいてクラスタリングし，その頻度分布を楽曲の音響的特徴を表現する音響ベクトルとして定義する (楽曲 j の音響ベクトルは $\mathbf{a}_j$ と表す．$\mathbf{a}_j$ の要素数は，ベクトル量子化におけるクラス数 L である)．

2.3 関連付けのための変換行列の推定方法

2.1 節，2.2 節で定義した文書ベクトル $\mathbf{t}_j$ と音響ベクトル $\mathbf{a}_j$ を以下のように，線形変換によって関連付けることを考える．

$$\mathbf{a}_j = W\mathbf{t}_j \quad (10)$$

ここで変換行列 W は，楽曲 j の音響ベクトル $\mathbf{a}_j$ と，W に文書ベクトル $\mathbf{t}_j$ をかけた $W\mathbf{t}_j$ との 2 乗誤差 $||\mathbf{a}_j - W\mathbf{t}_j||^2$ の楽曲の総数 J に関する総和が，最小となるように求める．

$$\hat{W} = \underset{W}{\operatorname{argmin}} \frac{1}{J}\sum_{j=1}^{J} ||\mathbf{a}_j - W\mathbf{t}_j||^2 \quad (11)$$

ここで推定する変換行列 W は正方行列とした．すなわち，文書ベクトル $\mathbf{t}_j$ の要素数 k と音響ベクトル $\mathbf{a}_j$ の要素数 L は等しいものとする．

3 評価実験

2 種類のテキストデータを利用して，提案する文書ベクトルと音響ベクトルの関連付け手法の性能について調査する．

3.1 楽曲を解説したレビューの利用

音楽ダウンロードサイトにおいて，楽曲を解説したレビューを利用して，文書ベクトルと音響ベクトルの線形変換による関連付け手法の性能を調査する．

3.1.1 使用データ

音楽ダウンロードサイト Mora[8] における試聴曲 (約 30 秒程度) と，その曲を解説したレビューを提案手法の学習と評価のために利用する．試聴曲とレビューは 1 対 1 の関係にあり，アルバム曲全体を解説したレビューは使用しない．その結果，合計 2,650 曲の音響信号（試聴曲）とレビューを収集した．レビューあたりの平均文章数は，2.74 文であった．茶筅 ver.2.3.3[9] を利用して形態素解析を行った結果，形態素の種類は 115,388 であった．そのうち品詞を名詞，動詞，形容詞に限定した場合，形態素の種類は 10,578 であり，これを文書ベクトル $\mathbf{x}_j$ の要素数 I とする．また，収集した音響信号に対して，フレーム長 32ms，フレームシフト 16ms で短時間スペクトルを求め，2.2 節で提案した楽曲の音響的特徴を表現するための音響ベクトルを計算した．スペクトルフラットネスの次数は 14 とした．

3.1.2 評価方法

2,650 曲の音響信号とレビューとのペアを 5 つのグループに分割する．そして，4 つを学習データ，残りの 1 つを評価データとして 5-fold クロスバリデーションを行う open データによる評価と，4 つを学習データ，そのうちの 1 つを評価データとして利用

した closed データによる評価を行う．学習データから推定された変換行列 W に，評価データである楽曲 m の文書ベクトル $\mathbf{t}_m$ をかけて推定される音響ベクトル $W\mathbf{t}_m$ と真の音響ベクトル $\mathbf{a}_m$ との 2 乗誤差 $e_{m,m} = ||\mathbf{a}_m - W\mathbf{t}_m||^2$ が ε 以内であれば正解とする．また，別の楽曲 l の音響ベクトル $\mathbf{a}_l$ と $W\mathbf{t}_m$ との 2 乗誤差 $e_{l,m} = ||\mathbf{a}_l - W\mathbf{t}_m||^2$ も利用して，変換行列 W の推定性能を評価するために，情報検索システムの評価に利用される再現率，適合率の考え方を取り入れる．再現率は，評価データの楽曲数に対して，正解と出力された楽曲数の割合であり，(12) 式のように定義する．

$$\text{再現率 } (R) = \frac{e_{m,m} \leq \varepsilon\text{をみたす楽曲数}}{\text{評価データの楽曲数}} \qquad (12)$$

一方，適合率は，評価データを入力したとき，2 乗誤差が ε 以内であった楽曲数に対して，どれだけ正解が含まれているかという正確性の指標として，(13) 式のように定義する．

$$\text{適合率 } (P) = \frac{e_{m,m} \leq \varepsilon\text{をみたす楽曲数}}{e_{l,m} \leq \varepsilon\text{をみたす楽曲数}} \qquad (13)$$

ε を変化させ，楽曲の音響ベクトルと文書ベクトルがどれだけ正確に，また網羅的に変換行列 W によって関連付けられているかについて検証する．

3.1.3 実験結果

変換行列 W のサイズを 1,024(音響ベクトルの要素数)×1,024(次元削減後の文書ベクトルの要素数) に固定し，ε を変化させたときの再現率と適合率を図 4 に示す．ε を大きくすると適合率は下降し，再現率は上昇する．また，closed データによる評価と比べて，open データによる評価で関連付け性能が低いことを確認した．この open データに適応できない原因の 1 つとして，変換行列 W の学習が十分でないことが考えられる．使用した 2,650 曲にさらに曲を追加して学習データ量と関連付け性能との関係について調査する必要がある．今回は楽曲に対して単一のレビューを使用したが，楽曲に対する複数のレビューを大量に集めること，歌詞等のテキストデータを加えることにより，文書ベクトル抽出のための学習データを増やす必要性も考えられる．また，提案した音響ベクトルによって楽曲の音響的特徴をとらえることが十分であるかについても検討する必要がある．

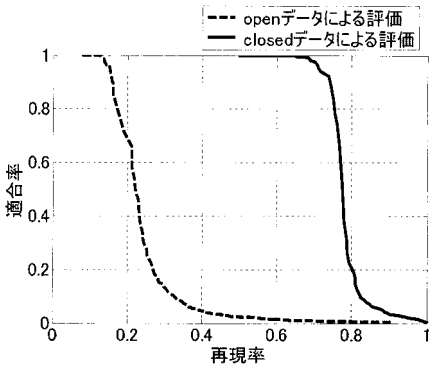

図 4　楽曲のレビューを用いた時の文書ベクトルと音響ベクトルの関連付け性能

3.2 曲名とアーティスト名を含む Web ページの利用

楽曲のレビューに限定せずに，曲名とアーティスト名を含む大量の Web ページを利用して，関連付け性能を評価する．

3.2.1 使用データ

3.1.1 節で使用した楽曲から 30 曲を選び，これらの曲名とアーティスト名を含む Web ページを収集する．検索エンジン Google の検索クエリとして曲名とアーティスト名を入力し，検索結果の上位 100 位以内の Web ページすべてを，楽曲を解説したテキストとして利用し，関連付け性能を評価する．収集した Web ページ (総数 2,176 個) に対して，茶筅 ver.2.3.3 を利用して形態素解析を行った結果，形態素の種類は 29,036 であった．そのうち品詞を名詞，動詞，形容詞に限定した場合，形態素の種類は 26,956 であり，これを文書ベクトル $\mathbf{x}_j$ の要素数 I とする．

3.2.2 評価方法

2,176 個の Web ページのうち 30 個（各曲に 1 つ）の Web ページを評価データとし，残りの Web ページを学習データとした．今回は評価データと学習データは重複しない open データで評価実験を行った．検索対象曲が 30 曲と少数であるため，性能の評価尺度には，N ベスト正解率を利用した．評価用の楽曲 m を解説した Web ページについて，変換行列に文書ベクトルをかけて得られる推定される音響ベクトル $\hat{\mathbf{a}}_m (= W\mathbf{t}_m)$ と，対応する曲の音響ベクトル $\mathbf{a}_m$ との

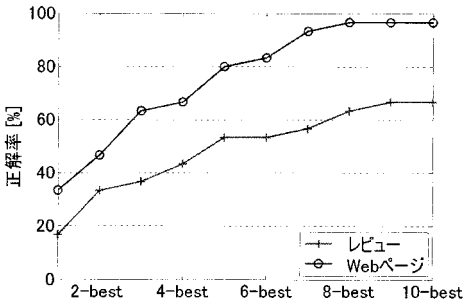

図 5　楽曲を解説したテキストとして Web ページを利用した時の，関連付け性能の評価:使用した楽曲数が 30 曲と少ないため，N ベスト正解率で評価した

2 乗誤差を $e_{m,m} = ||\mathbf{a}_m - \hat{\mathbf{a}}_m||^2$ とする．また，残りの検索対象 (29 曲) 中の曲 l の音響ベクトル $\mathbf{a}_l$ との 2 乗誤差を $e_{l,m} = ||\mathbf{a}_l - \hat{\mathbf{a}}_m||^2$ とする．そして，誤差の小さい順に並べた場合に $e_{m,m}$ が上位 N 位以内にあれば正解とする．

3.2.3　実験結果

3.1.3 節と同様に，変換行列 W のサイズを 1,024(音響ベクトルの要素数)×1,024(次元削減後の文書ベクトルの要素数) にして関連付けを行った結果を図 5 に示す．評価用のテキストを Web ページとし，変換行列の学習に Web ページ (30 曲 ×71.5 個の Web ページの計 2,146 個) を利用した場合と，レビュー (2,650 曲 ×1 個のレビューの計 2,650 個) を利用した場合との関連付け性能を比較する．学習と評価データのテキストの種類が同じ Web ページが，レビューを利用して変換行列を学習した場合より高い関連付け性能を確認した．この理由として，各楽曲に対し 1 つのレビューを利用するより，Web ページを複数利用することにより，楽曲を解説したテキストとしての語彙の偏りが解消される．つまり，大量のテキストを利用することにより，出現する語彙が楽曲を表現するにふさわしい語彙へと収束すると考えられる．また，Web ページを収集した際には，各楽曲に対してテキストを多数収集することを目的としたため，特に Web ページの種類（ブログによる曲の感想や掲示板での批評等）については考慮しなかった．Web ページの種類による関連付け性能への影響を調査し，関連付けの学習に有効なテキストについて検討する必要がある．

4　まとめと今後の課題

閲覧中の Web ページにふさわしい BGM を，自動的に選曲する楽曲検索（推薦）システム構築のために，楽曲を解説したテキストに出現する形態素を TF-IDF によって重みづけした文書ベクトルと，楽曲の音響的特徴を表現するために SFM 等の頻度分布を利用した音響ベクトルを提案し，線形変換で関連付けることを提案した．2,650 曲の試聴曲とレビューを利用し再現率，適合率による評価を行った．また，各楽曲に対し単一のレビューより曲名とアーティスト名を含む Web ページを複数利用すると関連付け性能が高いことを確認した．今後は，関連付けの学習のために各楽曲に対しどの程度のテキストデータ量が必要か，また，関連付けに有効なテキストの種類について調査する予定である．さらに，主観評価実験による性能評価も行う予定である．

参考文献

[1] 熊本忠彦，大田公子：印象に基づく楽曲検索システムの設計・構築・公開，人工知能学会論文誌，Vol. 21, No. 3, pp. 310–318 (2006).

[2] D. Turnbull, L. Barrington and G. Lanckriet: Modelling music and words using a multi-class naive bayes approach, *Proceedings of the International Symposium on Music Information Retrieval* (2006).

[3] B. Whitman and D. Ellis: Automatic record reviews, *Proceedings of the International Symposium on Music Information Retrieval* (2004).

[4] M. Slaney: Semantic-audio retrieval, *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (2002).

[5] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 1, pp. 391–407 (1990).

[6] G. Tzanetakis and P. Cook: Musical genre classification of audio signals, *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, pp. 293–302 (2002).

[7] G. Tzanetakis and P. Cook: MARSYAS: A framework for audio analysis, *Organized Sound*, Vol. 4, No. 3 (2000).

[8] Mora: http://mora.jp/.

[9] 松本祐治，北内　啓，山下達雄，平野善隆，松田　寛，高岡一馬，浅原正幸：日本語形態素解析システム『茶筌』Ver.　2.3.3，奈良先端科学技術大学院大学 (2003).