# 雰囲気を反映したBGM推薦システムの提案

坂井 栞[1,a)]　高屋 英知[1]　池田 圭佑[2]　川野 陽慈[1]　佐藤 圭[1]　山内 和樹[3]　大矢 隼士[3]　栗原 聡[1]

**概要**：我々は日常生活をおくる上で「空気を読むこと」について意識しないことはない．「空気」というものには実態がないが，日本人は度々物事の決定を「空気」に委ねることがある．よって場の「空気」である場の雰囲気をデザインするということが非常に重要になってくる．我々はその中で BGM に着目した．BGM には人をリラックスさせる感情誘導効果や店の雰囲気を明るくするイメージ誘導効果などが挙げられ，大きな労力を必要とすることなく空間の雰囲気を変えることが可能である．以上から本研究は店舗の雰囲気を反映した BGM 推薦するシステムの構築を行う．撮影された店舗内動画に異なる環境音を組み合わせ，仮想店舗動画を作成する．動画・楽曲それぞれにに対しラベル付けを行ったものを正解データとして学習を行う．店舗内動画との類似度が高いものを推薦し，店舗の動画に対して適切であるか評価実験を行った．

# Proposal of BGM Recommend System Reflecting Atmosphere

Sakai Shiori[1,a)]　Takaya Eichi[1]　Keisuke Ikeda[2]　Yoji Kawano[1]　Kei Sato[1]　Yamauchi Kazuki[3]
Hayato Oya[3]　Satoshi Kurihara[1]

## 1. はじめに

近年，KY（空気が読めない）という略語がメディアで取り上げられるように，我々は日常生活をおくる上で「空気を読むこと」について意識しないことはない．この「空気」という概念は日本独特のもので，日本人は度々物事の決定を「空気」に委ねることがある．その背景には民族性の違いが挙げられ, 日本人は限られた人間関係の中で上手くやり過ごすことを必要とされた．すなわち「空気」を読むということが求められたのである．また場の依存性[*1]が強く，存在する場によって考えや感情が左右される．このことからも人がその人らしくいられる「空気」を作る必要があり，場の雰囲気をデザインするということが非常に重要になってくる．

場の雰囲気を形成するものは，大きく分けて視覚情報と聴覚情報である．視覚情報は雰囲気をつくるという点においても大きな要素であり重要視されている．一方で聴覚情報，即ち場の音というものは，変更することは比較的容易であるにも関わらず，雰囲気のデザインでは比較的考慮されていないことが多い．そこで今回この雰囲気形成に寄与する音，その中でも BGM に着目した．

BGM には感情誘導効果やイメージ誘導効果を持つとされ [1]，大きな労力を必要とすることなく空間の雰囲気を変えることが可能である．現にレストランなどではそれらの効果を利用しようと楽曲配信サービスを導入しているところが多い．しかし時間帯による店内の見え方や客層の変化により，店舗側はその都度雰囲気にあった BGM を提供することが求められる．刻々と変化する店内の状況に BGM 配信サービスが対応することは非常に困難である．また店側で楽曲選択を行うにも膨大な曲数の中から合った曲を選出する労力と時間を捻出する必要が生まれる．

そこで，本研究は店舗の雰囲気に適合する BGM を推薦

[1] 電気通信大学 大学院情報理工学研究科
The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[2] 電気通信大学 大学院情報システム学研究科
The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[3] 株式会社レコチョク
RecoChoku Co.,Ltd., Shibuya, Tokyo 150–0002, Japan
[a)] ssakai@ni.is.uec.ac.jp
[*1] 対象物を知覚するときに，その背景や環境に影響を受けやすい性質のこと

するシステムの構築を行う．今回はあらかじめ撮影された店舗内動画と異なる環境音を組み合わせ，印象評価を行い，類似度が高い楽曲を推薦する実験および評価を行う．

## 2. 関連研究

画像や楽曲といったコンテンツに対して，ユーザがタグ付けすることはソーシャルタギングと呼ばれ，それらを利用した推薦や検索システムの構築については多く研究されている．

梶ら [2] らは歌詞とアノテーションを利用し，視聴時のユーザの状況に合わせたプレイリストを作成するために楽曲とユーザを特徴量空間へマップする手法を採用している．特徴量は歌詞，楽曲情景，視聴状況を用いており，それらの特徴量空間にユーザをマップすることで，楽曲間，ユーザと楽曲間，またユーザ間の類似度計算を可能にしている．楽曲情景のラベルは登場人物（一人，私），いつ（朝，過去，春），状況（恋愛中，反社会），心理状況（悲しい，怒り）の 4 項目を用いている．歌詞と楽曲情景については．それまで視聴した好きな曲の特徴量平均をそれぞれの特徴量空間にそのユーザの嗜好としてマップすることで．推薦を行っている．Kaminskas[3] らの研究ではユーザが関心のある場所（place of interest, POI）に即した楽曲推薦のシステムを構築している．楽曲と POI に双方同様の感情語を用いたタグをつけ，それらをベクトルとして扱い，類似度から適した楽曲を推薦している．タグには 9 項目の感情タグと 13 項目の物理的タグ（色や気温など）を用いている．

## 3. 本研究のアプローチ

本研究は店舗における BGM 利用を想定して，場の雰囲気に合った BGM 推薦システムを提案する．本章では BGM 推薦システムの概要と新たな印象評価ラベルを提案する．

### 3.1 本研究の提案

#### 3.1.1 環境音の利用

昨今の研究において，場所に対してラベル付けを行っているものが多くあり，その多くが画像に対して行われている．しかし容易にデータを得られるが一方で，視覚情報しか得られないという問題がある．そこで本研究では動画を印象評価に用い，その中でも環境音に着目する．

環境音は我々が普段生活する中で意識しなくても耳にし，音によってどのような場所であるのかを判断する．また環境音によって場所の見え方が変化するという傾向がある [6]．本研究では BGM を独立した音として扱うのではなく，環境音を含めて場をデザインすることを考慮し，環境音を含めた印象評価を行う．

#### 3.1.2 新規ラベルの作成

近年，感情語や印象語を用いて画像や楽曲にラベル付けを行い，楽曲などを推薦する研究が多く存在する．ラベルを用いることで画像や楽曲がユーザーへ与える心理的影響を考慮することができ，個人の嗜好や状況に即した楽曲などを推薦することが可能である．しかし本研究は BGM の推薦であり，個人の経験や感情に依存する感情語や印象語ラベルを使用することは不特定多数への適応を目的とした BGM 推薦には不向きである．

そこで本研究では，「情景」の項目においてスターバックスや東急ハンズといった具体的な店舗名を用いてラベル付けを行う．チェーン店は全体を通してコンセプトを持っており，店舗内装などを統一しているところが多い．そのため，店を利用した人の間では同一のイメージを共有することが可能である．また今までは複数の感情語や印象語を用いることで店舗内を表現していたが，具体的な店名を用いることで一つのラベルで表現することが可能になる．以上から具体的店名のラベル付けで情報の抜け落ちを防ぎ，より正確な雰囲気の評価を目指す．

### 3.2 システムの概要

本研究における場の雰囲気を反映した BGM 推薦の流れを図 1 に示す．

まず，店舗内動画像を提案システムに入力し，店舗内画像の特徴ベクトルを抽出する．同様に使用したい楽曲群を入力し，システム内で楽曲をメル周波数ケプストラム係数及びスペクトログラムに変換して楽曲特徴ベクトルを抽出する．それぞれの楽曲と店舗内動画の類似度を算出し，類似度が高いものを店舗に適した楽曲として推薦する．

## 4. データセット

### 4.1 ラベルの選定

本研究では情景ラベルにおいて実際の店舗名を使用する．ラベルを決めるにあたり，予備実験を行った．

回答者は学生 4 人，社会人 2 名で，店舗イメージが確立している店舗名を列記してもらった．集計後，店の種類に偏りが出ないよう，USEN のコンシェルジュサービス [7] を参考に記載されている項目で補った．ラベル内容は表 1 に示す．これらのラベルを店舗内動画と楽曲評価のときに

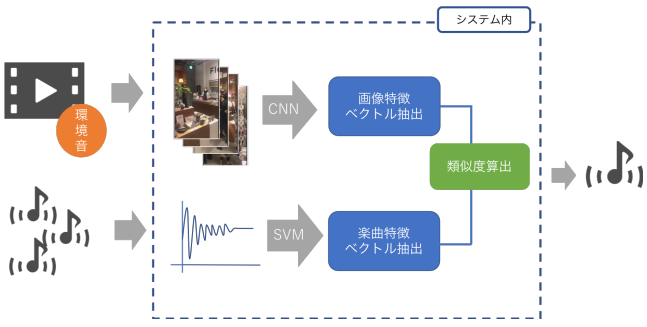


図 1 システムの概要図

図 2　店舗内動画イメージ図

使用した．

### 4.2　店舗動画のデータセット作成

店舗内の雰囲気を評価するための店舗内動画を収集した．

撮影は目線の高さで店舗内を 180 度撮影とし．時間は 10 秒から 15 秒ほどで行った．画像サイズは $1920 \times 1080$，フレーム数は 30fps とする．撮影イメージを図 2 に示す．ライトやアプリなどで明度や彩度の調整はしないものとする．

収集された店舗内動画から環境音を切り離し，異なる 5 種類の環境音を付け加えることで仮想店舗内動画を作成した．環境音はカフェ店内，子供が多いイベント会場内，ショッピングモール内，オフィス内，街中を使用した．

表 1　ラベル一覧

| 情景 | |
|---|---|
| スターバックス | ダイソー |
| ルノアール | 東急ハンズ |
| コメダ珈琲 | 無印良品 |
| バー | ニトリ |
| 居酒屋 | 大塚家具 |
| 割烹・料亭 | 紀伊國屋書店 |
| ラーメン屋 | TSUTAYA |
| 沖縄料理店 | 高島屋 |
| イタリア料理店 | PARCO |
| 西友 | イオンモール |
| 成城石井 | シェラトンホテル |
| カルディ | オフィス |
| アパレル（高級店） | 企業ロビー（大企業） |
| アパレル（フォーマル） | 企業ロビー（中小企業） |
| アパレル（カジュアル） | 国際空港 |
| ドラッグストア | 地方空港 |
| francfranc | |
| 日差し | |
| あり | なし |
| 時間帯 | |
| 朝（-10 時台） | 昼（11 時台-14 時台） |
| 夕方（15 時台-17 時台） | 夜（18 時台-） |
| 都市度合い | |
| 都会 | 郊外 |
| 田舎 | |

### 4.3　印象評価

#### 4.3.1　店舗内動画評価

仮想店舗内動画に対して印象評価を行ってもらった．作成した仮想店舗内動画を視聴してもらい，動画がどのような状況に当てはまるか，表 1 のラベルから選択してもらった．

#### 4.3.2　楽曲評価

楽曲を聴取し，どのような状況下で BGM として流れているかということを基準に評価を行ってもらった．楽曲は「J-POP」，「アニメ」，「キッズ・ファミリー」，「歌謡曲・演歌」，「邦楽ヒップホップ・R & B・レゲエ」，「邦楽ロック」，「洋楽ヒップホップ・R & B・レゲエ」，「洋楽ポップス」，「洋楽ロック」，「洋楽総合」の計 10 個のジャンルから 100 曲ずつを用意した．楽曲がどのような状況下に当てはまるか，表 1 のラベルから選択してもらった．

## 5.　識別器の構築と推薦方法

### 5.1　楽曲特徴量抽出

本研究ではメル周波数ケプストラム係数を用いて，楽曲識別器を構築した．

メル周波数ケプストラム（Mel Frequency Cepstral Coefficient，以下 MFCC）は，フーリエ変換によって求めたスペクトル情報に対して，低い周波数では細かく，高い周波数では荒い分解能を持つ人間の聴覚特性に合わせたフィルタを通して，その出力を対数変換し，さらにこれを離散コサイン変換したものである．

人間は低い周波数においては少しの高さの違いでも感じ取れるが，高い周波数の音はある程度高さが変わらないと変化したように感じない．人間が感じる音の高さの変化を一定にしたものがメル尺度と呼ばれるものである．メル尺度上で一定間隔になるようにパワースペクトルのベクトルのデータを計測し，さらに近傍のデータとの平均化操作を行うフィルタバンク処理を行うことで，人間の感じる周波数情報に近いものが得られる．人間の近くは対数スケールであるため，メル帯域スペクトルを対数化する．このとき，周波数成分には音源情報だけでなく声道情報が混在するためパワースペクトルを変換してこの 2 つの成分を線形和に置き換え，フィルタリングにより両者を分離する．

$v(n)$ を時刻 $n$ の声門波，$h(n)$ を時刻 $n$ の声道のインパルス応答とし，フーリエ変換すると $Y(n)$，$V(n)$，$H(n)$ はそれぞれ $y(n)$，$v(n)$，$h(n)$ となる．音声のパワースペクトル $S(k)$ は，

$$S(k) = |V(k)|^2 |H(k)|^2$$

となる．両辺の対数をとると，

$$\log S(k) = 2 \log |V(k)| + 2 \log |H(k)|$$

となる．さらにこのパワースペクトルの対数に対し逆フー

リエ変換を適用する．

$$
\begin{aligned}
c(n) &= \frac{1}{N}\sum_{k=0}^{N-1}\log S_k \exp(j\frac{2\pi kn}{N}) \\
&= \frac{1}{N}\sum_{k=0}^{N-1}\log S_k cos(\frac{2\pi kn}{N}) \\
&= \frac{2}{N}\sum_{k=0}^{N-1}\log V_k cos(\frac{2\pi kn}{N}) + \frac{1}{N}\sum_{k=0}^{N-1}\log H_k cos(\frac{2\pi kn}{N})
\end{aligned}
$$

以上よりケプストラム $c(n)$ を得ることができる．横軸にケフレンシー[*2]，縦軸にケプストラムの値をとる．スペクトル包絡の成分 $H_k$ は低ケフレンシー領域に現れ，声門波は高ケフレンシー領域に出るため，リフタリングを行い，低ケフレンシー成分のみを取り出したものが MFCC として音声認識などに利用される．

### 5.2 識別器構築

本研究では店舗内動画と楽曲の識別を行う．その際，店舗内動画は CNN を用いて学習を行い，楽曲に関しては SVM と CNN を並行して識別の構築も試みた．その後，楽曲のラベルを出力し，比較を行った．

#### 5.2.1 サポートベクターマシン（SVM）

サポートベクターマシン（Support Vector Machine，以下 SVM）とは法則に関係ありそうな要素を特徴ベクトルによって表し，データに潜む複雑な法則性の発見を最適化問題に帰着して効率的に解くことができるアルゴリズムである．

特徴量が多くなっても精度が良く，比較的少数のデータでも良い結果になりやすい，パラメータの算出が容易である．しかしその一方で訓練データの数が増えると計算に時間がかかる．固定長のベクトルに限らず配列データや木構造やグラフ構造に対しても 2 つのデータの間の関係を数値化する専用の計算式さえ設計すれば学習を効率的に行うことができる．

本研究での学習の設定を表 5.2.1 に示す．

表 2　SVM における識別器の設定

| カーネル | RBF カーネル |
|---|---|
| コストパラメータ $C$ | 100 |
| RBF カーネルパラメー $\gamma$ | 0.00001 |

#### 5.2.2 畳み込みニューラルネットワーク（CNN）

本節では，店舗内動画と楽曲の識別に使用する畳み込みニューラルネットワーク (Convolutional Neural Network，以下 CNN) について解説する．CNN は，人の顔の認識や道路標識の認識など，画像認識に特化したネットワークである．本研究における店舗識別用 CNN の内部構造は図 3，楽曲分類用 CNN の内部構造は図??の通りである．

[*2] 時間と同じ次元で，周波数 (frequency) から作られた造語

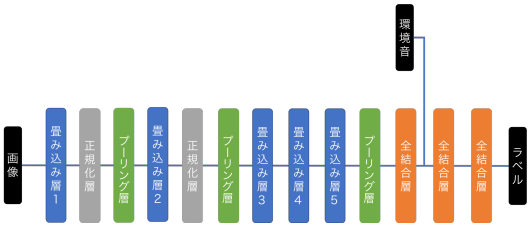


図 3　店舗識別用 CNN のアーキテクチャ

入力画像として画像を与え，畳み込み層により画面の特徴を抽出する．畳み込み層では，入力画像をフィルターで畳み込み，画像をぼかしたり，エッジ (色が変化する境目) を強調するなど，画像の特徴を捉える．プーリング層では縦・横方向の空間を小さくするような処理を行い，正規化層では各画像に固有の明るさ加減やコントラストの強さを整える．店舗識別用の CNN においては，全結合層で環境音のベクトルとの結合を行う．そして出力されたデータと正解データを用いて平均二乗誤差を計算し，最小になるように誤差伝播法により重みを更新する．

### 5.3 楽曲ラベル出力結果比較

CNN におけるラベル出力結果では全ての楽曲においてほぼ同様のラベルが出力され，一方で SVM は楽曲ごとに異なるラベル値が出力された．またあらかじめ付けられたラベルと出力されたラベル間で距離を計算したところ平均 20.78 と CNN の平均 80.73 よりも遙かに小さい値が算出された．

以上から，本研究では楽曲のラベル出力に SVM で構築した識別器を利用する．

### 5.4 楽曲推薦方法

識別器から得られる結果は，各ラベルにおける値である．動画と楽曲のラベルの類似度を計算し，値が大きい 3 曲を推薦する．本研究ではユークリッド距離，コサイン類似度，ピアソン相関係数を用いて求める．式は以下の通りである．

$X = \{x_1, x_2, ..., x_n\}$，$Y = \{y_1, y_2, ..., y_n\}$，$\overline{X} = \frac{x_1+x_2+...+x_n}{n}$，$\overline{Y} = \frac{y_1+y_2+...+y_n}{n}$ としたとき，

- コサイン類似度

$$CosineSim(X,Y) = \frac{x_1y_1 + x_2y_2 + ... + x_ny_n}{\sqrt{{x_1}^2 + ... + {x_n}^2}\sqrt{{y_1}^2 + ... + {y_n}^2}}$$

- ピアソン確率相関関数

$$PiasonSim(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- ユークリッド距離

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$$

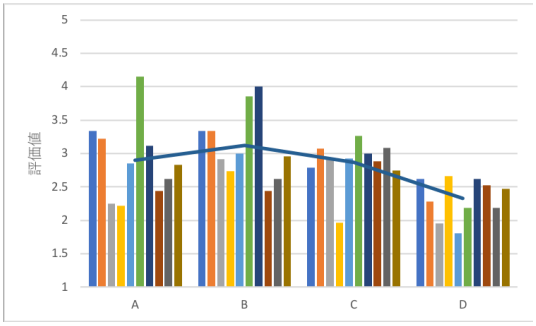


**表 3** 推薦楽曲結果

$$EuclidSim(X,Y) = \frac{1}{d+1}$$

## 6. 評価実験

本評価実験は，被験者に動画を視聴してもらい評価アンケートを通してシステムを評価する．

1 本の動画に対し 3 種類の類似度計算を用い，3 曲ずつ出力した．それらの店舗内動画を視聴しながら推薦された楽曲を試聴し，5 段階で評価をしてもらった．被験者には本学から募集した学生で，1 本の動画に対し 8 名または 9 名の被験者が評価を行った．

結果を図 6 に示す．A はコサイン類似度，B がピアソン相関係数，C がユークリッド距離を用いて類似度計算を行ったグループである．

グループごとの平均値が 2.99，3.12，2.89 の値であった．A，C と比べて B の平均値は高いものの類似度計算により評価が大きく異なることはなかった．またグループ内の標準偏差は 0.59，0.46，0.32 であった．図からも分かるように A と B においては一部突出して評価が高いものが存在し，C は点数にばらつきが見られなかった．

## 7. 評価実験考察

評価アンケートを通し，どの類似度計算においても最適な BGM 推薦には至らなかった．原因の一つとしてデータの少なさとラベルの多さが学習に影響を与えたと考えられる．また本研究は複数のラベルを用いており，基本的に 2 クラスの識別で用いられている SVM には不向きの分類であったことが挙げられる．

また十分なラベル付けがされていないという問題もあった．本研究ではデータセット作成時に動画や楽曲に対してラベル付けをしてもらったが，複数のラベルがつけられたものと，そうでないものの差が激しかった．楽曲に対するタグ付けは 1 曲に対し最低 2 人で行っていたことでラベル付与の偏りが生まれ，推薦に影響が出たのではないかと考えている．

加えて評価実験において同じ楽曲であるにもかかわらず，グループが異なるだけで点数が異なるものが多く存在した．前の曲との関係性が原因と考えられるが，確証を得るためにも人為的に様々な曲と組み合わせ，評価を行う必要がある．その上で店舗動画と楽曲の類似度だけでなく，楽曲間の類似度・相性を考慮する必要がある．

## 8. おわりに

本研究は店舗での利用を想定し，店舗内での雰囲気を反映した BGM 推薦システムを提案した．環境音を含めて店舗内の印象評価を行い，また不特定多数の人の間で共通認識を利用するために具体的な店名をラベルに使用した．店舗内動画と楽曲それぞれで学習を行い，3 種類の類似度計算を用いて動画に対し楽曲を推薦した．他の類似度計算と比べるとピアソン相関係数によって推薦された楽曲が BGM として適しているという結果が出たが，適切な推薦には至らなかった．

今後の課題はデータセットの拡充と楽曲間の相性を考慮したシステムの提案が必要となってくる．

**謝辞** 楽曲は株式会社レコチョクに提供していただきました．

### 参考文献

[1] D. Västfjäll, *Emotion induction through music: A review of the musical mood induction procedure[Special issue 2001-2002]*, Musicae Scientiae, pp.171-203, 2002.
[2] 梶克彦，平田圭二，長尾確，状況と嗜好に関するアノテーションに基づくオンライン楽曲推薦システム，情報処理学会研究報告，Vol.2004，pp.33-38，2004.
[3] M. Kaminskas and F. Ricci, *Location-adapted music recommendation using tags*, Adaption and Personalization，pp.183-194，2011.
[4] T. Zhang and J. Kuo, *Audio content analysis for online audiovisual data segmentation and classification*, Trans. Speech Audio Processing, vol. 9, pp. 441-457, May 2001.
[5] L. Mion and G. D. Poli, *Score-independent audio features for description of music expression*, IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 2, pp. 458 - 466, 2008.
[6] R Murray Schafer, *The Soundscape, Our Sonic Environment and the Tuning of the World*, Destiny Books, 1976.
[7] music.usen.com，コンシェルジュ．http://music.usen.com（閲覧日：2017 年 12 月 5 日）