体験映像に対する音インデクスについての分析
Analysis about Sound Index for Experience Movies

志村 将吾 † Shogo Shimura　**平野 靖** ‡ Yasushi Hirano　**梶田 将司** ‡ Shoji Kajita　**間瀬 健二** ‡ Kenji Mase

1. はじめに

情報処理機器の小型化，ハードディスクの記録容量の増加といった背景から，著者らは，体験を常時記録する研究を行っている [1]．体験を常時記録することで，一瞬の出来事を逃すことなく記録可能である．しかし，長時間記録された映像の中には，ユーザにとって意味の無いシーンが多く含まれており，全てを閲覧することは効率が悪い．したがって，ユーザにとって重要な体験を抽出する必要がある．

そこで，重要な体験をしていると感じた場合には，ユーザ自身がインデクスの付与を行い，そのインデクスを基に，類似画像検索を用いることで体験区間の推定をすることを考える．体験記録を行うユーザの負担を考えたとき，ユーザは記録の際に必要最小限であるカメラとマイクロフォン以外の機器を身に付けることは望ましくない．ユーザの体の一部を用いて発せられる音をインデクスとして使用すれば，特別な機器を用いることなく，マイクロフォンによって記録することができる．体の一部を使用して発生させられる音の典型的なものとして，手を叩く音と指を鳴らす音がある．日常生活において，これらの音の抽出が可能であることを実験により確認した．

2. 音の特徴分析

インデクスとして用いる手を叩く音と指を鳴らす音の 2 種類について，その周波数構造を分析した．

まず，手を叩く音の周波数構造を図 1 に示す．0-2kHz の間に，大きなピークが 1 つあることがわかる．ここで図中の $Power$ は，FFT 変換後の各周波数成分の実部と虚部を用いて，

$$Power = 10 \times \log_{10}\left((\text{実部})^2 + (\text{虚部})^2\right) \quad (1)$$

で表される．

指を鳴らす音の周波数構造を図 2 に示す．2-4kHz の間に，大きなピークが 1 つあることがわかる．

3. 抽出方法の検討

3.1 手を叩く音

まず，0-2kHz 区間に大きなピークが存在することから，この区間の平均パワーが大きくなることが予想される．また，人間の音声や車が近くを通るときのノイズは，0-2kHz 区間に最も多く存在することが，事前の実験により確認された (図 3)．これらの音は，0-2kHz 区間にいくつもの鋭いピークが存在し，手を叩く音に比べて各ピークの幅は小さい．さらに分散は，図 3 のようなノイズであれば小さく，図 1 のように大きいピークが存在すれば大きくなると予想される．したがって，0-2kHz 区間の平均パワー，ピークの幅，パワーの分散を特徴量として採用した．

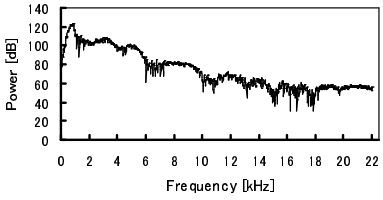


図 1: 手を叩く音の周波数構造

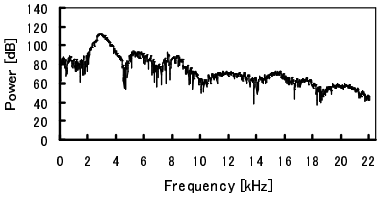
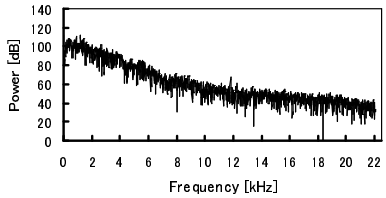
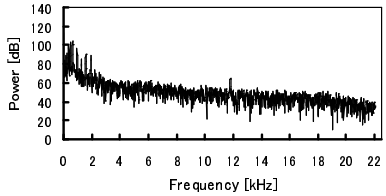


図 2: 指を鳴らす音の周波数構造



(a) 車が近くを通る音の例



(b) 人間の音声の例

図 3: ノイズの周波数構造

3.2 指を鳴らす音

まず，2-4kHz 区間に大きなピークが存在することから，この区間の平均パワーが大きくなることが予想される．また，図 3(a) から，0-2kHz 区間だけでなく，2-4kHz 区間におけるノイズの影響も無視することはできない．しかし，指を鳴らす音とノイズを比較してみると，前者の音は 2-4kHz 区間の面積が突出している．したがって，2-4kHz 区間の平均パワー，全体の面積に対してこの区間の面積が占める割合を特徴量として採用した．

4. 実験

4.1 予備実験

前述した特徴量を用いて，「手を叩く音」，「指を鳴らす音」の 2 種類の音が，どの程度の精度で抽出可能であるか検証する実験を行った．静かな室内，静かな屋外，

†名古屋大学大学院情報科学研究科社会システム情報学専攻
‡名古屋大学情報連携基盤センター

静かな室内　　静かな屋外　　騒音の多い屋外

図 4: 実験環境

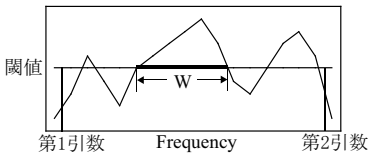


図 5: 関数 $maxwidth$

騒音の多い屋外の 3 箇所 (図 4) で，2 種類の音を直立した状態で 20 回ずつ発生させ，胸部に装着したマイクロフォンを用いて，サンプリングレート 44.1kHz，分解能 16bit で収録した．

収録されたデータに対して，サンプル点数 4096 個で FFT を行い，以下の式により特徴量を計算した．式 (2)〜(4) は，0-2kHz 区間の平均パワー，ピークの幅，パワーの分散を求める式で，式 (5), (6) は，2-4kHz 区間の平均パワー，全体の面積に対してこの区間の面積が占める割合を求める式である．式中の $P(i)$ は振幅スペクトルであり，FFT 変換後の i 番目の係数を $F(i)$ としたとき，$\sqrt{(F(i)\text{ の実部})^2 + (F(i)\text{ の虚部})^2}$ である．また，n_1 は，$n \times (44100/4096) < 2000$ を満たす最大の整数 n，n_2 は，$n \times (44100/4096) < 4000$ を満たす最大の整数 n である (本実験の場合は，$n_1 = 185, n_2 = 371$)．なお，式 (3) の関数 $maxwidth$ は，第 1 及び第 2 引数で区間を指定し，第 3 引数を閾値として，閾値より大きい値である区間の幅のうち最大のものを返す関数である (図 5)．また，P_{max} は，0-2kHz 区間で最大の $P(i)$ である．

$$\left(\begin{smallmatrix}0\text{-}2\text{kHz の}\\ \text{平均パワー}\end{smallmatrix}\right)\quad A_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} P(i) \tag{2}$$

$$(\text{ピークの幅})\quad W = maxwidth(1, n_1, P_{max}/2) \tag{3}$$

$$(\text{パワーの分散})\quad V = \frac{1}{n_1}\sum_{i=1}^{n_1}(A_1 - P(i))^2 \tag{4}$$

$$\left(\begin{smallmatrix}2\text{-}4\text{kHz の}\\ \text{平均パワー}\end{smallmatrix}\right)\quad A_2 = \frac{1}{n_2 - n_1}\sum_{i=n_1+1}^{n_2} P(i) \tag{5}$$

$$(\text{面積比})\quad R = \sum_{i=n_1+1}^{n_2} P(i) / \sum_{i=1}^{2048} P(i) \tag{6}$$

以上の特徴量を用いて音を抽出した結果を表 1 に示す．なお，各特徴量に対する閾値は適合率が最大になるように決定した．

手を叩く音については，再現率は 90%以上，適合率は

表 1: 抽出結果 1

(a) 手を叩く音

	静かな室内	静かな屋外	騒音の多い屋外
再現率	90%	95%	100%
適合率	100%	100%	100%

(b) 指を鳴らす音

	静かな室内	静かな屋外	騒音の多い屋外
再現率	100%	100%	100%
適合率	100%	100%	100%

表 2: 抽出結果 2

	手を叩く音	指を鳴らす音
再現率	100%	100%
適合率	67%	100%

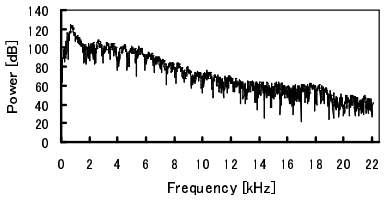


図 6: 靴を高いところから落とした音の周波数構造

すべての環境で 100%であった．また，指を鳴らす音については，全ての環境で再現率，適合率が共に 100%であった．

4.2 実環境実験

続いて，実際の生活を考慮した場合では，どの程度の抽出精度が得られるか確認する実験を行った．胸部にマイクロフォンを装着し，図 4 に示した 3 箇所を歩き回った．その際，各環境で 2 種類の音を無作為に 2 回ずつ，計 6 回の音を発生させ，4.1 に示したサンプリングレート，及び分解能で収録した．

収録されたデータに対して，4.1 と同じ特徴量を用いて音を抽出した結果を表 2 に示す．なお，閾値は 4.1 と同じ値を用いた．

手を叩く音の適合率は 67%であったものの，他の項目では 100%となった．実環境における実験で，検索の際，抽出漏れの無いことが確認された．

手を叩く音と誤って抽出された音は，靴を高いところから落とすときに生じる音であった．この音の周波数構造を図 6 に示す．0-2kHz の区間に大きいピークがあり，図 1 の手を叩く音と構造が類似している．そのため，誤って抽出したと考えられる．

5. おわりに

本稿では，体験記録におけるインデクスとして，ユーザの体の一部を使用して発生させられる音に着目し，手を叩く音，指を鳴らす音の 2 種類について分析した．実験を行い，これらの音が，再現率は 90%以上，適合率は 67%以上の精度で抽出可能であることが確認された．特に，指を鳴らす音については，全ての実験で再現率，適合率が共に 100%であった．

今後の課題として，本稿での提案手法を用いて，実際に体験区間の推定を行うことが挙げられる．

謝辞

本研究は文部科学省「知的資産の電子的な保存・活用を支援するソフトウェア基盤技術の構築」プロジェクトの支援により行われた．

文献

[1] Shogo Shimura, Yasushi Hirano, Shoji Kajita and Kenji Mase, “Experiment of Recalling Emotions in Wearable Experience Recordings”, Advances in Pervasive Computing: Adjunct Proceedings of the Third International Conference on Pervasive Computing, pp. 19-22 (2005).