語感の定量化に向けた 画像に写る概念に対する視覚的印象の推定

松平 茅隼[1,a)]　カストナー マークアウレル[2,1,b)]　駒水 孝裕[1,c)]　平山 高嗣[3,1,d)]　井手 一郎[1,e)]

概要

我々は語感の定量化を目指している．そのために，まず 1 段階目として，語の発音に対する画像を生成し，次に 2 段階目として，画像に写る概念に対する視覚的印象を推定する枠組を考えている．本発表では，この 2 段階目を実現するために，事前学習済みモデルである CLIP，GPT，Stable Diffusion を用いて，形容詞対で記述することができる任意の視覚的印象の推定手法を提案する．実際に，4 種の視覚的印象を対象として各印象を示す概念が写る画像を収集し，各印象に対応する概念が写る画像の検索タスクとして評価実験を実施した．その結果，画像に写る概念に対する各印象の度合いをほぼ正確に推定できることを確認した．

1. はじめに

語感とは，人間が語の発音から連想し得る意味を指す．語感を引き起こす要因の 1 つに音象徴性 [6], [10] がある．例えば，大半の人間にとって，母国語によらず「マルマ」よりも「タケテ」という発音の方が，視覚的に尖った印象を連想させる発音であることが知られている [6]．一方で，言語話者の語彙知識に依存する言語固有の要因も考えられる．英語において ‘gl-’ から始まる単語の多くが光に関連した意味を持つことから，接頭辞 ‘gl-’ が光を意味することを主張する Phonestheme [4] がその例である．

本研究ではこれらの要因のうち，後者に着目し，語の発音から連想される概念が持つ視覚的印象の定量化を目指す．これを実現するため，本研究では図 1 に示す 2 段階の枠組を提案する．第 1 段階では，任意の語の発音が入力されると，まず発音エンコーダが発音特徴を算出し，それを基に画像生成モデル [9] が画像を生成する．我々の先行研究では，既存語を用いて調音音声学に基づく発音類似度の制約 [5] を導入した発音エンコーダを学習することで，存在しない語の発音が入力された際に，発音が類似した既存語（類音語）のそれに類似した発音特徴を算出する手法とともに，その発音特徴を画像生成モデルと統合し，未知の語の発音に対してその類音語の視覚的概念を写す画像を生成する手法を構築した [7]．第 2 段階では，その生成画像に対する視覚的印象の推定を行なうことで，語の発音が持つ視覚的印象の定量化を行なう．本発表では，この第 2 段階の手法に関して報告する．

[1] 名古屋大学
[2] 広島市立大学
[3] 人間環境大学
[a)] matsuhirac@cs.is.i.nagoya-u.ac.jp
[b)] mkastner@hiroshima-cu.ac.jp
[c)] taka-coma@acm.org
[d)] t-hirayama@uhe.ac.jp
[e)] ide@i.nagoya-u.ac.jp

2. 語感の定量化に関する関連研究

語の発音から連想される印象の定量化を試みた既存研究は存在するが，言語によらず適用可能でかつ一般的で客観的な手法は未だ存在しない．黒川は主に日本語を対象として，音素の調音的特徴に着目した主観的見解に基づく語感の分析法を提案した [11], [12] が，客観的な評価を行なっていない．小松・秋山は対象を日本語のオノマトペに限定し，上記を含む音象徴性に対する主観的見解を統合して印象推定システムを提案した [13] が，日本語固有の知見に基づくため，他の言語へは適用できない．近年，大規模なデータで事前学習されたモデルが人間の感覚と相関がある “語感” を出力できることが明らかになり [1], [3]，その利用が検討されている．しかし，既存研究は特定の視覚的印象（例：語感の鋭さ／丸さ）の推定可能性に対する検証に留まっており，任意の印象に対する汎用性は未検証である．

本研究では，まず語の発音に対して画像生成を行なうことで，語から連想される概念の視覚的特徴を可視化する．次に，その画像に対して視覚的印象を推定することにより，形容詞対を用いて記述することができる任意の視覚的印象に対応した語感の定量化を目指す．

3. 画像に写る概念に対する視覚的印象の推定

本研究では，事前学習済みの Vision-Language モデル CLIP [8]，言語生成モデル GPT [2]，Text-to-Image 生成モデル Stable Diffusion [9] を用いて，画像に写る概念に対する任意の視覚的印象を推定する手法を提案する．ここで，CLIP と Stable Diffusion を用いて生成画像中に写る

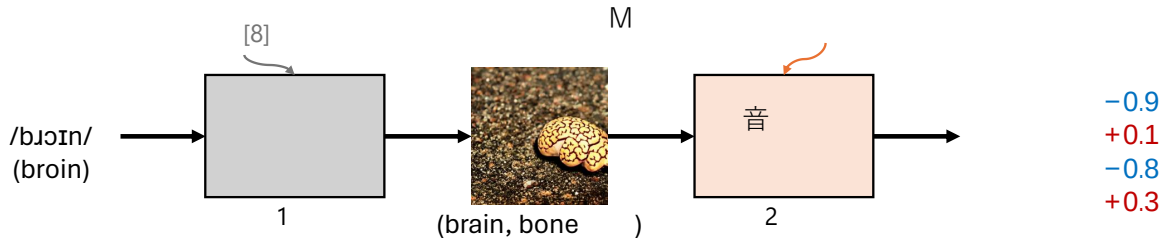


図 1: 語の発音から想起される概念の視覚的印象を定量化する 2 段階の枠組.

表 1: GPT に与えるプロンプト．[POSITIVITY] 及び [NEGATIVITY] は対象とする視覚的印象を表す名詞で置換される．

```
Pick up 10 adjectives describing the [POSITIVITY] of an object.
Then, as antonyms, please pick up another 10 adjectives describing the [NEGATIVITY] of an object.\n
During the word selection, please be careful about the following five points:\n
- All words must be commonly used adjectives.  Also, they must not be nouns.\n
- They must not contain hyphens.\n
- Their meanings must be clear.  They must not be interpreted to mean other things.\n
- Their word stems must be diverse.  That is, no two words must have the same stem.\n
- They can be used in the attributive position.  That is, they can be used in the noun phrase
"the <WORD> thing" where <WORD> represents each selected adjective.\n
```

物体の視覚的な鋭さ／丸さを推定した先行研究が存在する [1]．彼らはまず形状の鋭さ／丸さを形容する形容詞集合対を選定し，その集合対に含まれる各語に対して Stable Diffusion [9] を用いて画像生成を行ない，生成画像集合対を得た．次に，それらを CLIP の特徴空間上に写像し，生成画像集合対間の識別境界に直交する Probe Vector と呼ばれるベクトルを算出することで，視覚的な鋭さ／丸さを推定する手法を提案した．本研究では，GPT を用いてこれを任意の視覚的印象へ拡張する手法を提案する．

3.1 形容詞集合対の選定

まず，対象となる視覚的印象に関して，対義関係をなす形容詞集合の対 $(\mathrm{A}_+, \mathrm{A}_-)$ を定義する．先行研究 [1] では，形状の鋭さ／丸さに対する形容詞を 10 語ずつ選定しているが，選定基準が明確でない．そこで本研究では，GPT を用いて[*1] これを自動的に選定する手法を提案する．

対象とする視覚的印象に関する対義語の組（例：鋭さ／丸さ）が与えられた際，本手法では表 1 に示すプロンプトを GPT に入力する．これにより，対象とする視覚的印象について正負それぞれの度合いを形容する形容詞が 10 語出力される．ここで，GPT の出力は確率的であることから，この結果は試行の度に変化しうる．GPT の出力において最も尤度が高い語を選定するため，本研究では上記プロンプトの入力を 50 回試行し，最も多い回数出力された上位 10 語を A_+ 及び A_- として採用する．この際，プロンプトに記述した指示に沿わない語（例：ハイフンが含まれる）が含まれる場合，それを除外した上で上位 10 語を選定する．また，プロンプトとして入力する視覚的印象を表す名詞（例：sharpness）の派生元となる形容詞（例：sharp）を A_+ と A_- それぞれに必ず含めるため，その語が上位 10 語に含まれなかった場合には，その語と上位 9 語の計 10 語を採用する．

3.2 Probe Vector を用いた視覚的印象推定値の算出

前項で選定した形容詞集合対を用いて，先行研究が提案する Probe Vector [1] を算出する．そのために，まず A_+ 及び A_- に含まれる各形容詞 a に対して画像を生成し，生成画像集合対 $(\mathrm{I}_{\mathrm{A}_+}, \mathrm{I}_{\mathrm{A}_-})$ を得る．本研究では，Stable Diffusion[*2] を用いて，各形容詞 $a \in \mathrm{A}_+ \cup \mathrm{A}_-$ に対する画像を 50 枚生成する．この際，画像生成のプロンプトとして ‘a photo of a <形容詞> object’ を指定する．次に，CLIP の画像エンコーダを用いて a に対する生成画像 50 枚から特徴ベクトルを算出し，それらの平均値を形容詞 a に対する特徴ベクトル $\boldsymbol{w}_a$ とする．

次に，この特徴ベクトル $\boldsymbol{w}_a$ を用いて，形容詞集合対に対する Probe Vector $\boldsymbol{w}_{\mathrm{adj}}$ を算出する．具体的に，Probe Vector $\boldsymbol{w}_{\mathrm{adj}}$ は式 (1) で与えられる．ただし，$\hat{\boldsymbol{w}}_a$ は大きさが 1 になるよう正規化されたベクトル $\frac{\boldsymbol{w}_a}{\|\boldsymbol{w}_a\|_2}$ を指す．

$$\boldsymbol{w}_{\mathrm{adj}} = \sum_{a \in \mathrm{A}_+} \hat{\boldsymbol{w}}_a - \sum_{a \in \mathrm{A}_-} \hat{\boldsymbol{w}}_a \qquad (1)$$

本式により，2 クラス $\{\hat{\boldsymbol{w}}_a \mid a \in \mathrm{A}_+\}$ 及び $\{\hat{\boldsymbol{w}}_a \mid a \in \mathrm{A}_-\}$ を識別する方向を向くベクトル $\boldsymbol{w}_{\mathrm{adj}}$ が得られる [1]．

最後に，ある概念が写る入力画像 p に対して，この Probe Vector を用いて視覚的印象の度合い $\gamma_p \in [-1, 1]$ を算出する．これは，p に対する CLIP 画像特徴ベクトル $\boldsymbol{v}_p$ と $\boldsymbol{w}_{\mathrm{adj}}$ との間の余弦類似度 $\gamma_p = \hat{\boldsymbol{v}}_p \cdot \hat{\boldsymbol{w}}_{\mathrm{adj}}$ として表される．Probe

[*1] 本発表では一貫して gpt-3.5-turbo を用いる．

[*2] 本発表では一貫して stable-diffusion-v1.4 を用いる．

表 2: 本実験で対象とした視覚的印象，GPT に入力するプロンプト上で置換する名詞，及び選定された形容詞集合対.

印象	プロンプト中の名詞	GPT により選定された形容詞集合
鋭さ	[POSITIVITY]= sharpness	A_+={sharp, piercing, jagged, pointed, cutting, serrated, keen, acute, edgy, prickly}
丸さ	[NEGATIVITY]= roundness	A_-={round, spherical, circular, smooth, curved, oval, bulbous, rotund, blunt, plump}
大きさ	[POSITIVITY]= largeness	A_+={large, massive, immense, enormous, gigantic, huge, vast, colossal, mammoth, tremendous}
小ささ	[NEGATIVITY]= smallness	A_-={small, tiny, petite, miniature, diminutive, microscopic, lilliputian, compact, minuscule, mini}
速さ	[POSITIVITY]= quickness	A_+={quick, swift, speedy, brisk, hasty, rapid, fast, fleet, prompt, agile}
遅さ	[NEGATIVITY]= slowness	A_-={slow, leisurely, gradual, languid, tardy, sluggish, plodding, delayed, lethargic, ponderous}
柔らかさ	[POSITIVITY]= softness	A_+={soft, tender, smooth, fluffy, silky, pliable, gentle, delicate, velvety, plush}
硬さ	[NEGATIVITY]= firmness	A_-={firm, tough, rigid, solid, resilient, robust, hard, sturdy, steady, strong}

表 3: 各視覚的印象に対して GPT を用いて選定した，正負それぞれの印象を想起させうる概念 10 語の集合.

印象	プロンプト中の名詞	GPT により選定された概念 10 語の集合
鋭さ	[POSITIVITY]= sharpness	{needle, sword, razor, dagger, knife, scissors, axe, arrow, saw, chisel}
丸さ	[NEGATIVITY]= roundness	{ball, orange, globe, peach, sphere, melon, circle, plum, cherry, bubble}
大きさ	[POSITIVITY]= largeness	{mountain, whale, elephant, building, canyon, ship, glacier, sequoia, airplane, galaxy}
小ささ	[NEGATIVITY]= smallness	{ant, atom, pebble, dust, grain, molecule, insect, microbe, seed, drop}
速さ	[POSITIVITY]= quickness	{cheetah, rocket, bullet, ferrari, lightning, speedboat, jet, falcon, arrow, greyhound}
遅さ	[NEGATIVITY]= slowness	{sloth, snail, tortoise, glacier, slug, loris, koala, turtle, iceberg, panda}
柔らかさ	[POSITIVITY]= softness	{marshmallow, cloud, pillow, cotton, feather, silk, sponge, blanket, velvet, mousse}
硬さ	[NEGATIVITY]= firmness	{brick, concrete, steel, rock, wood, plastic, glass, marble, diamond, metal}

Vector の性質より，γ_p が 1 に近いほど画像 p が A_+ に含まれる形容詞が含意する視覚的印象を示し，かつ A_- に含まれる形容詞が含意する視覚的印象を示さないといえる.

4. 視覚的印象推定手法に対する評価実験

4.1 対象とする視覚的印象及び比較手法

本実験で対象とする視覚的印象と 3.1 項の手法により得られた形容詞集合対 (A_+, A_-) を表 2 に示す.

3.1 項で提案した GPT を用いた形容詞集合対の選定に対する比較手法として，形容詞集合対を人手で選定して評価した．鋭さに関しては先行研究 [1] のものを使用し，他の印象に関しては GPT による単語選定の実施前に著者ら自身が選定した.

4.2 評価用画像データの収集

評価用画像データとして，対象となる視覚的印象を確実に示す概念が写る画像が好ましい．そこで，GPT と Bing 画像検索[*3]及び Stalbe Diffusion を併用することで，そのような実画像及び生成画像を収集した.

まず，3.1 項と同様に，GPT に名詞を 10 語出力させるプロンプトを作成し，そのプロンプトを GPT に 50 回与え，頻出した名詞上位 10 語を得た．これにより，対象とする各視覚的印象について，正負それぞれの印象を想起させうる概念を 10 語ずつ得た．各印象に対して実際に選定された概念 10 語ずつを表 3 に示す．なお，既存の企業名との重複などが原因で Bing 画像検索結果が対象とする視覚的印象を示していないと考えられた単語（例：“apple”—丸さ，“thunderbolt”—速さ）は，人手で除外した.

次に，各視覚的印象に対するこれら 20 語それぞれに対し，実画像及び生成画像を収集した．実画像を得る場合は Bing 画像検索を利用し，各語に対して画像検索結果の上位 20 画像を収集した．これにより，対象とする視覚的印象それぞれについて，正の印象を想起させうる概念 10 語に対する画像 $10 \times 20 = 200$ 枚，負の印象を想起させうる概念 10 語に対する画像 200 枚を得た．また，生成画像を得る場合は，各語に対して Stable Diffusion を用いて画像を 50 枚生成することで，同様に正・負の印象についてそれぞれ $10 \times 50 = 500$ 枚の画像を得た．ここで，画像生成のプロンプトには ‘a photo of a <概念>’ を指定した.

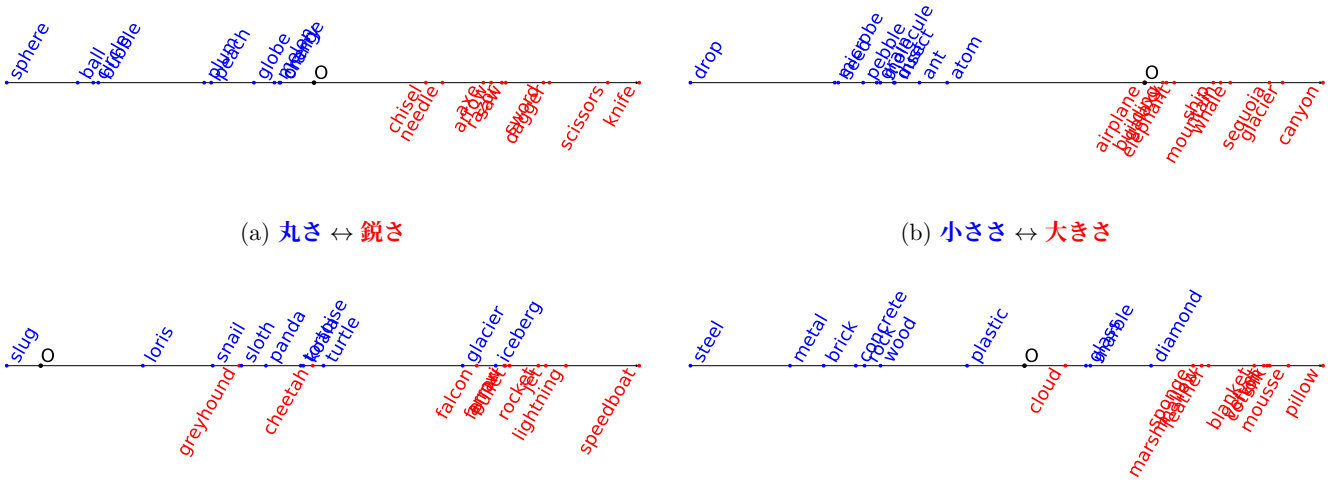
4.3 タスク及び評価指標

各視覚的印象に対する上記評価用画像データを用い，関連画像の検索タスクとして提案手法の評価を行なった．つまり，各印象に対して正・負に関連する計 400 枚の実画像集合または計 1,000 枚の生成画像集合から，正に関連する画像 200 枚または 500 枚を正例として検索するタスクとする．検索対象となる正例の集合が既知であるため，提案手法による印象推定値が高い画像上位 k 枚を検索した際の再現率 Recall@k（k は 200 または 500）を評価指標とした.

[*3] https://www.bing.com/images/feed（2024/5/20 参照）

表 4: 提案手法による視覚的印象推定値を用いた画像検索タスクに関する実験結果.

	実画像 & Recall@200（全画像：400 枚）				生成画像 & Recall@500（全画像：1,000 枚）			
形容詞選定手法	鋭さ／丸さ	大きさ／小ささ	速さ／遅さ	柔らかさ／硬さ	鋭さ／丸さ	大きさ／小ささ	速さ／遅さ	柔らかさ／硬さ
GPT	**0.990**	**0.955**	0.780	**0.830**	**0.956**	**0.976**	0.784	**0.852**
人手（鋭さ／丸さ [1]）	0.910	0.950	**0.795**	0.825	0.894	0.974	**0.794**	0.828



(a) 丸さ ↔ 鋭さ　(b) 小ささ ↔ 大きさ

(c) 遅さ ↔ 速さ　(d) 硬さ ↔ 柔らかさ

図 2: 評価実験で使用した各視覚的印象を示す単語それぞれの生成画像 50 枚に対する印象推定値 γ_p の平均値の分布. 数直線上でより右側に存在する語ほど印象推定値の平均値が高い．また，数直線上の O は原点を示す.

4.4 実験結果

実験結果を表 4 に示す．まず，提案手法が実・生成画像ともに高い再現率を示したことから，実・生成画像いずれの場合でも，画像に写る概念の視覚的印象の度合いを正負混同せず推定できていることがわかる．また同表より，GPT を用いた単語選定が人手によるそれに劣らないことも確認できる．特に「速さ／遅さ」以外の視覚的印象については，人手による単語選定での結果を上回る再現率が得られた．これにより，画像に写る概念に対する視覚的印象推定における GPT を用いた単語選定の有効性を確認した.

定性的な評価に向けて，本実験で使用した各視覚的印象を示す概念 10 語それぞれについて，その生成画像 50 枚に対する印象推定値の平均値の分布を図 2 に示す. 本図から，評価指標にも現れているように，提案手法は鋭さや大きさに比べ，速さや柔らかさの正確な推定に失敗していることがわかる．例えば図 2(d) から，提案手法が “diamond” や “glass” に対する生成画像を硬くないと誤って推定していることが見られる．これは，提案手法における GPT を用いた形容詞の選定過程で，硬さに対する形容詞集合 A_ に透明でかつ硬い物体を形容する形容詞が含まれなかったことが原因であると考えられる．これに対処するためには，より多様な視覚的印象を形容する形容詞の選定手法が求められる．また図 2(c) では，提案手法が “greyhound” や “cheetah” に対する生成画像を速くないと推定し，一方で “glacier” や “iceberg” に対する生成画像を遅くないと誤って推定していることが見られる．速さ／遅さに関する印象はそもそも画像単体からでは推定が困難であることが原因であると想定されるため，これらをより正確に推定するためには，例えば動画像生成技術の利用などが考えられる.

5. まとめ

本研究では語感の定量化を目指し，本発表では事前学習済みモデルである CLIP，GPT，Stable Diffusion を用いて画像中に写る概念が示す任意の視覚的印象を推定する手法を提案した．4 種の視覚的印象に対して，各印象の推定値に基づく画像検索タスクとして実施した評価実験の結果，画像に写る概念に対する各印象の度合いを正負混同せずほぼ正確に推定できることを確認した.

今後の課題として，本枠組全体に対する評価，より抽象的な視覚的特徴への拡張などが挙げられる.

謝辞

本研究の一部は，科研費（22H03612），MSR-CORE16 プログラムによる．本研究は、JST 次世代研究者挑戦的研究プログラム JPMJSP2125 による「東海国立大学機構メイク・ニュー・スタンダード次世代研究事業」の財政支援を受けたものである。本研究は名古屋大学のスーパーコンピュータ「不老」の一般利用を利用して実施した.

参考文献

[1] Alper, M. and Averbuch-Elor, H.: Kiki or Bouba? Sound symbolism in vision-and-language models, *Adv. Neural Inf. Process. Syst.*, Vol. 36, pp. 78347–78359 (2023).

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D.: Language models are few-shot learners, *Adv. Neural Inf. Process. Syst.*, Vol. 33, pp. 1877–1901 (2020).

[3] Cai, Z. G., Duan, X., Haslett, D. A., Wang, S. and Pickering, M. J.: Do large language models resemble humans in language use?, *Comput. Res. Reposit., arXiv Preprints*, arXiv:2303.08014 (2024).

[4] Firth, J. R.: *The Tongues of Men and Speech*, Oxford University Press, London, England, UK (1964).

[5] International Phonetic Association: *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press, Cambridge, England, UK (1999).

[6] Köhler, W.: *Gestalt Psychology*, H. Liveright, New York, NY, USA (1929).

[7] Matsuhira, C., Kastner, M. A., Komamizu, T., Hirayama, T., Doman, K., Kawanishi, Y. and Ide, I.: Interpolating the text-to-image correspondence based on phonetic and phonological similarities for nonword-to-image generation, *IEEE Access*, Vol. 12, pp. 41299–41316 (2024).

[8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning transferable visual models from natural language supervision, *Proc. 38th Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.*, Vol. 139, pp. 8748–8763 (2021).

[9] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-resolution image synthesis with latent diffusion models, *Proc. 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 10684–10695 (2022).

[10] Sapir, E.: A study in phonetic symbolism, *J. Exp. Psychol.*, Vol. 12, No. 3, pp. 225–239 (1929).

[11] **黒川伊保子：怪獣の名はなぜガギグゲゴなのか，新潮社，東京，日本**（2004）.

[12] **黒川伊保子：商標評価手法の一考察—ことばの感性評価，知財管理，** Vol. 56, No. 5, pp. 745–752（2006）.

[13] **小松孝徳，秋山広美：ユーザの直感的表現を支援するオノマトペ意図理解システム，電子情報通信学会論文誌**（A），Vol. J92-A, No. 11, pp. 752–763（2009）.