2004－DBS－134　(I)　(23)
2004／7／13

音声メディアデータを対象としたメタデータ自動抽出方式に関する研究

及川聡子[†1]　中西崇文[†2]　北川高嗣[†2]

本稿では，音声メディアデータを対象として，メタデータを自動抽出する方式について示す．具体的には，Media-lexicon　Transformation Operator ML の実装方式を示す．まず，Banse と Scherer の研究を用い，音声を分析して 29 個の音声パラメータを得る．つぎに，音声パラメータと印象語との相関から，重み付き印象語を求める．得られた重み付き印象語がメタデータであり，人間が音声から受ける印象を示す．

The System of Automatic Extraction System for Voice Media Data

SATOKO OIKAWA,[†1]　TAKAFUMI NAKANISSHI,[†2]

and　TAKASHI KITAGAWA[†2]

This paper presents the system which extracts metadata automatically for voice media data. Specifically, the mounting system of Media-lexicon Transformation Operator ML is shown. First, using research of Banse and Scherer, a sound is analyzed and 29 voice parameters are obtained. Next, correlation with a voice parameter and an impression word is asked for an impression word with weight. The obtained weight is metadata and the impression which man receives from a sound is shown.

1. はじめに

現在，コンピュータネットワーク上には多種多様なメディアデータ群が散在している．さらに，メディアデータ群から情報を獲得する機会が増加している．それらを検索対象とするシステムの実現が行われつつある．メディアデータ群を対象とした情報獲得の機会の可能性が増大する一方，適切な情報獲得方式の実現が重要な課題となっている．特に，これらの多種多様なメディアデータ群を扱うシステムにおいて，人間の感性は重要な問題のうちの 1 つとなっている．

メディアデータを対象とした検索方式はメディアデータの特徴量を直接比較することによって検索を行う直接的な方法と，メディアデータに付与された抽象データ(以下, メタデータ)を用いて検索する間接的な方法に大別できる．

我々は，メディアデータに対応するメタデータを言葉によって表現し，検索者の与える文脈に応じた意味的解釈を伴う間接的な検索方式として，メディアデータを対象とした意味的連想検索方式[1,2,3]を提案している．これにより，統計的に意味素を抽出して意味の解釈を実現する従来の研究[4]と比較して，言葉の意味を文脈に応じて解釈する機構より，言葉と言葉，あるいは，言葉とメディアデータ間の意味的な関係を与えられた文脈や状況に応じて動的に計算することが可能となる．

また我々は，文献[2,5,6,7]でメディアデータのメタデータを自動抽出するための実現方式について示している．特に文献[5]では，メディアデータの印象を表す語をメタデータとして自動抽出する枠組みとして，Media-lexicon Transformation Operator を示している．これは，メディアデータからその分野の専門家による研究や評論，統計などによる人間がそのメディアデータから受ける印象を表す単語を抽出する

†1 筑波大学大学院理工学研究科
Graduate School of Science and Engineering, University of Tsukuba
†2 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

Media-lexicon Transformation Operator を実現している．Media-lexicon Transformation Operator により，抽出された単語をメタデータとして，意味的連想検索を実現することにより，検索者が発行する印象語から，その印象に合致したメディアデータの検索が可能となる．

さらに我々は，文献[8]で異種メディア間において検索をおこなう異種メディア間連想検索を基本構造している．異種メディア間検索とは，例えば，画像メディアデータからその印象にあった楽曲メディアデータを意味的連想検索するような方式を指す．多数の異種のメディアデータが散在しているマルチメディア環境においては，異なるメディアデータ間の検索，統合によって新しい情報の生成が重要である．また，これらから，既存のデータの利用機会を増大させ，データベース群の利用価値を飛躍的に増大させることが可能となると考えられる．文献[8]の方式では，各異種メディア間を語と語を相関を計量することを可能とする意味の数学モデルによる意味的連想検索機構により計量しているため，メディアデータごとに Media-lexicon Transformation Operator を実現することにより，システム全体を変えることなく様々な異種メディア間検索の実現を可能としている．

一方，これまでのコンピュータと人間の論理情報の伝達だけでは，操作による人間の負荷が大きくなっており，人間の感性や直感に合致したユーザへの負担が少ない，コミュニケーションメディアの実現が重要な課題となってきている．一般に，我々のコミュニケーションにおいて，互いの感情を正確に理解することが重要であり，メディアデータ群を対象とした環境において，感性的な行動を人間とシステムとのコミュニケーションメディアとして導入されれば，人間の感性や直感に合致した，ユーザへの負荷が少ないインターフェースの実現が可能となると考えられる．様々なメディアデータ，人間の行動，環境などからその印象を取り出す Media-lexicon Transformation Operator を基本機能として実現し，文献[8]の方式を適用することにより，人間の感性や直感に合致したユーザの負担が少ないインターフェースがこれまでのシステムを有効に活用しつつ実現ができると考えられる．

本稿では，人間が発する音声について注目する．人間は言葉を発することによってコミュニケーションを実現している．その言葉自身の意味だけではなく，声の調子など非言語的な特徴に感情が含まれていることが多い．そのため，人間が発した音声を取り出し，その音声が表す感情によって，様々なメディアデータを操作することによって，人間とシステムのコミュニケーションメディアとして導入が可能となると考えられる．

本稿では，音声メディアデータを対象とした Media-lexicon Transformation Operator について示す．これは，入力した音声と合致した印象を表す語をメタデータとして抽出する方式である．音声に関する研究として Banse と Scherer の研究がある．Banse と Scherer の研究は，音声から抽出できる 29 個の音声パラメータと，音声から人が受ける印象を表す語との関係を示している．本方式は Banse と Scherer の研究の関係を用いて，音声メディアデータからその印象を表す語をメタデータとして抽出する方式を示す．

2. 意味の数学モデルによる意味的連想検索方式の概要

本節では，意味の数学モデル[1]の概要を示す．人間が様々な印象を表す際に用いられる単語(以下, 印象語)によって表現した問い合わせに対応したメディアデータを検索することを目的とした意味の数学モデルによるメディアデータを対象とした意味的連想検索方式の概要を示す．詳細は，文献[1]に述べられている．

(1)メタデータ空間 MDS の設定
検索対象となるメディアデータをベクトルで表現したデータをマッピングするための正規直交空間(以下，メタデータ空間 MDS)を設定する．具体的な手順を以下に示す．
はじめに，m 個の基本データについて各々 n 個の特徴（$f_1, f_2, \cdots, f_n$）を列挙した特徴付きベクトル $\boldsymbol{d}_i (i = 1, \cdots, m)$ が与えられているものとし，そのベクトルを並べて構成する $m \times n$ 行列を M とおく（図 1 参照）．但し，M は列ごとに 2 ノルムで正規化されている．
(a)データ行列 M の相関行列 M^TM を計算する．
(b)M^TM を固有値分解する．

$$M^TM = Q \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_\nu & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} Q^T, \quad 0 \le \nu \le n. \tag{1.1}$$

ここで，行列 Q は，

$$Q = (\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n), \tag{1.2}$$

である．但し，q_i（i = 1，…，n）は，相関行列の正規化された固有ベクトルである．相関行列の対称性から，固有値はすべて実数であり，対応する固有ベクトルは互いに直交している．

(c)メタデータ空間 MDS を以下で定義する．

非ゼロ固有値に対応する固有ベクトルによって形成される正規直交空間をメタデータ空間MDS と定義する．同空間の次元 n は，データ行列 M のランクと一致し，n 次元ユークリッド空間となる．

$$\mathrm{MDS} := \mathrm{span}(q_1, q_2, \cdots, q_n). \quad (1.3)$$

$\{q_1, q_2, \cdots, q_n\}$ は，MDS の正規直交基底である．

(2)メタデータをメタデータ空間 MDS へ写像

設定されたメタデータ空間 MDS へメディアデータのメタデータをベクトル化し写像する．これにより，検索対象データのメタデータが同じメタデータ空間上に配置されることになり，検索対象データ間の意味的な関係を空間上でのノルムとして計算することが可能となる．

1.メディアデータの特徴付け

メディアデータ P を t 個の印象語（あるいは，t 個のオブジェクト）w_1，w_2，…，w_t を用いて，次のように特徴付けする．

$$P = \{w_1, w_2, \cdots, w_t\}. \quad (1.4)$$

ここで，各印象語 w_i は，データ行列の特徴と同一の特徴を成分とする特徴付きベクトルである．

$$w_i = (f_1, f_2, \cdots, f_n). \quad (1.5)$$

2.メディアデータベクトル P のベクトル表現

メディアデータベクトル p を構成する t 個の印象語 w_1，w_2，…，w_t が，それぞれ n 次元のベクトルで定義されている．印象語 w_1，w_2，…，w_t は，合成することで n 次元ベクトル表現され，メディアデータベクトル p を形成し，メタデータ空間 MDS に写像される．つまり,同じ空間上に言葉とメディアデータが配置される．したがって，言葉とメディアデータの関係を空間上の距離として動的に計算することが可能となる．

(3)メタデータ空間 MDS の部分空間(意味空間)の選択

検索者は与える文脈を複数の単語を用いて表現する．検索者が与える単語の集合をコンテクストと呼ぶ．このコンテクストを用いてメタデータ空間 MDS に各コンテクストに対応するベクトルを写像する．これらのベクトルは,メタデータ空間MDS において合成され，意味重心を表すベクトルが生成される．意味重心から各軸への射影値を相関とし，閾値を超えた相関値(以下，重み)を持つ軸からなる部分空間(以下，意味空間)が選択される．

(4)メタデータ空間 MDS の意味空間における相関の定量化

選択されたメタデータ空間MDS の部分空間(意味空間)において，メディアデータベクトルのノルムを検索語列との相関として計量する．これにより，与えられた

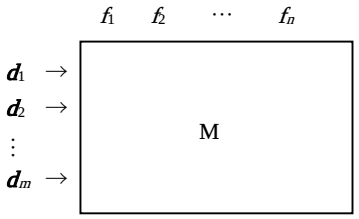

図 1　データ行列 M によるメタデータの表現.

Fig. 1. Representation of metadata items by matrix M.

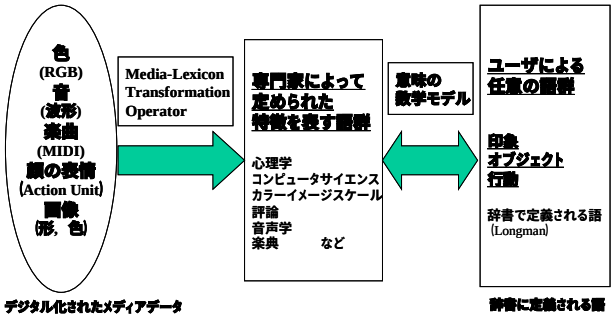

図 2　Media-lexicon Transformation Operator ML の概要.

Fig. 2.　Media-lexicon Transformation Operator ML.

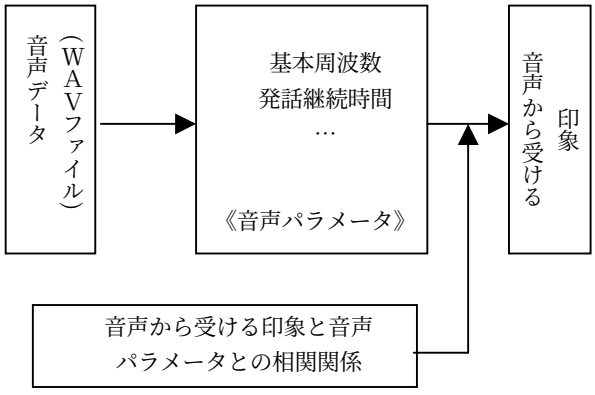

図 3　音声メタデータ生成方法.

Fig. 3. The generation method of voice metadata.

コンテクストと各メディアデータとの相関の強さを定量化している．この意味空間における検索結果は，各メディアデータを相関の強さについてソートしたリストとして与えられる．

3. 音声メディアデータを対象としたメタデータ自動抽出方式

本節では，デジタル化された音声情報を対象として，人間が音声から受ける印象を抽出し，言葉によって表現されるメタデータを自動的に抽出する音声メタデータ自動抽出方式について述べる．

まず, 音声から基本周波数, エネルギー(声の大きさ), スピーチ率，発声持続時間スペクトル，無声持続時間スペクトルなどの音声の構造を決定する要素（以下，音声パラメータ）を取り出す．次に，音声パラメータと音声の印象との相関を求める．ここで，音声パラメータと音声の印象との相関量を調べた Banse と Scherer の研究[3]を用いる．最後に，印象語の重みを計算する．重みは，相関量に近いほど強くなる．得られた重み付き印象語がメタデータである．

具体的には， Media-lexicon Transformation Operator ML[2]を実装する（図 2 参照）． 以後，音声情報はデジタル化され, WAV ファイルで与えられるものとする．

3.1 Media-lexicon Transformation Operator ML の概要

Media-lexicon Transformation Operator ML は，その分野の専門家による研究や評論, 統計などを用い，メディアデータから人間の受ける印象を, 単語で表し，メディアデータのメタデータとして抽出する．

本節では，Media-lexicon Transformation Operator ML の概要を述べ，図 3 に示す．

6 Step1：専門家の研究より，特徴量と印象語との相関を調べる

メディアデータから抽出可能な特徴量（色彩情報，テンポ，音程など）と印象語の関係を把握する．

Step2：メディアデータから特徴量ベクトル $\boldsymbol{p}$ を抽出する

メディアデータを解析し, メディアデータの印象をあらわす特徴量を抽出する．

Step3：特徴量ベクトル $\boldsymbol{p}$ から印象語を抽出する

特徴づけられた印象語と特徴量より, メタデータを抽出する．

3.2 Banse と Scherer の研究

Banse と Scherer の研究では，音声から取り出した 29 個の音声パラメータと 14 個の印象語群との相関を調べている．感情表現に声は重要な役割を果たす．しかしながら，音声には 2 つの有名な問題が存在している． 1つは，聞き手は声から感情を推測することができるか否かである．もう 1 つは，特定の感情に対応する特定の声のパターンがあるか否かである．

Banse と Scherer の研究は，プロの役者が感情をこめてシナリオを読み上げた音声をテストデータとして用いている．音声から取り出した 29 個の音声パラメータと 14 個の印象語群との相関を調べ，妥当性を確かめている．テストデータから，性別，感情表現の得手不得手を，正規化して取り除き，z 変換を施した結果が表 1 である. 但し, 印象語群は表 2 に示すとおり，聞き手が受け取る感情である．

したがって，Banse と Scherer が用いた方法で音声を分析し，取り出した音声パラメータから得られる値と表 1 の値との比較を行えば，重みつきの印象語，つまりメタデータを得られる．

3.3 メタデータ自動抽出方式

(1) 音声パラメータと印象語との関係

14 種類の印象語群を 29 個の音声パラメータの平均および標準偏差で特徴付けしたベクトル $\boldsymbol{w}_k$ を以下で定義する．

$$\boldsymbol{w}_k=(m_{k1}, m_{k2}, \cdots, m_{k29}, sd_{k1}, sd_{k2}, \cdots, sd_{k29})^{\mathrm{T}}. \quad (k = 1, 2, \cdots, 14) \tag{1.6}$$

但し，$\boldsymbol{w}_k$ の成分は， Banse と Scherer の相関表の数

表 1. 音声パラメータと印象語群との相関.

Table 1. Correlation with voice parameter and impression words.

Acoustic variable		c_1	c_2	…	c_{14}
fundamental frequency	M	1.1	0.16	…	-1.03
	SD	0.6	0.72	…	0.44
25 percentile	M	0.9	0.15	…	-0.93
	SD	0.7	0.73	…	0.31
…	…	…	…	…	…
Unvoiced long-term average spectrum 5000-8000Hz	M	-0	-0.3	…	0.21
	SD	0.3	0.46	…	1.09

表 2 印象語群.

Table 2. Impression words.

c_1	c_2	c_3	c_4	c_5	c_6	c_7
hot anger	cold anger	panic fear	anxiety	despair	sadness	elation
c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}
happiness	interest	boredom	shame	pride	disgust	contempt

値で，印象語 c_k に対応する音声パラメータの平均値，標準偏差がそれぞれ $m_{ki}(i = 1,\cdots,29)$ ， $sd_{kj}(j = 1,\cdots,29)$ に対応している．

(2) 音声パラメータの抽出

音声メディアデータから 29 の音声パラメータを抽出し，音声メディアデータベクトル $\boldsymbol{p}$ を定義する．

$$\boldsymbol{p} = (p_1, p_2, \cdots, p_{29})^T, \tag{1.7}$$

(3) 音声パラメータからメタデータを抽出

音声パラメータと印象語群との関係を表す w_{ki}，および抽出した音声メディアデータベクトル $\boldsymbol{p}$ より，音声メタデータベクトル I を抽出する．音声メタデータベクトル I を以下に示す．

$$I = (s_1, s_2, \cdots, s_{14}). \tag{1.8}$$

ここで，s_k を次式で定義する．

$$s_k = \frac{\alpha}{\sum_{i=1}^{29} \frac{\beta(m_{ki} - p_i)^2}{|sd_{ki}|} + \varepsilon}, (k = 1, 2, \ldots, 14) \tag{1.9}$$

但し，εは s_k の発散を防ぐ微小値であり，α，βはそれぞれ実験で定まる係数である．

(1.3)から，s_k は，音声から取り出した音声パラメータの値が，表1に近いほど大きな値をとることがわかる．つまり，s_k は，音声が各印象語群とどれほど似通っているかを示す重みである．

以上より，重み付き印象語群，つまり，メタデータが求められる．

4. おわりに

本稿では，音声メディアデータを対象としたメタデータ抽出手法を示した．今後の課題は，Media-lexicon Transformation Operator ML[2]を実装し，抽出したメタデータを異種メディア間検索へ用い，有効性を確認することである．

参考文献

1) Kitagawa, T. and Kiyoki, Y.: The mathematical model of meaning and its application to multidatabase systems, Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp. 130-135(1993).

2) Kiyoki, Y., Kitagawa, T. and Hayama, T.: A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning, Multimedia Data Management-- using metadata to integrate and apply digital media --, McGrawHill, A. Sheth and W. Klas(editors), Chapter 7 (1998).

3) 清木康,金子昌史,北川高嗣: 意味の数学モデルによる画像データベース探索方式とその学習機構, 電子情報通信学会論文誌 ,D-II,Vol.J79-D-II,No. 4,pp. 509-519 (1996).

4) Michael, W. B., Susan, T. D., Gavin, W. O.: Using linear algebra for intelligent information retrieval,SIAM Review Vol. 37, No.4, pp.573-595 (1995)..

5) T. Kitagawa, Y. Kiyoki, “Fundamental framework for media data retrieval systems using medialexco transformation operator,” Information Modeling and Knowledge Bases, Vol.12, pp. 316-326, 2001.

6) 北川高嗣, 中西崇文, 清木康: 楽曲メディアデータを対象としたメタデータ自動抽出方式の実現とその意味的楽曲検索への適用, 電子情報通信学会論文誌 Vol.J85-D-I,No.6,pp.512-526, 2002.

7) 北川高嗣, 中西崇文, 清木康: 静止画像メディアデータを対象としたメタデータ自動抽出方式の実現とその意味的画像検索への適用, 情報処理学会論文誌: データベース, VOl.43,No.SIG12(TOD16), pp38-51, 2002.

8) Takafumi Nakanishi, Takashi Kitagawa, Yasushi Kiyoki,: An Implementation Method of Associative Search for Heterogenous Mediadata Utilizing the Mathematical Model of Meaning and its Application to Image Data and Facial Expression, 2003 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM '03), pp.613-618, August(2003).

9) Banse. R. and Scherer, K, ”Acoustic profiles in vocal emotion expression,” Journal of Personality and Social Psychology, 70, pp.614-636, 1996.