特集論文

環境音を対象とした擬音語自動認識

擬音語表現における音素決定曖昧性の解消

Sound-Imitation Word Recognition for Environmental Sounds

Disambiguation in Determining Phonemes of Sound-Imitation Words

石原 一志 Kazushi Ishihara
京都大学情報学研究科
Graduate School of Informatics, Kyoto University
ishihara@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~ishihara/

駒谷 和範 Kazunori Komatani
（同　上）
komatani@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~komatani/

尾形 哲也 Tetsuya Ogata
（同　上）
ogata@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~ogata/

奥乃 博 Hiroshi G. Okuno
（同　上）
okuno@kuis.kyoto-u.ac.jp, http://winnie.kuis.kyoto-u.ac.jp/~okuno/

keywords: sound-imitation word, onomatopoeia, retrieval system, environmental sounds

Summary

Environmental sounds are very helpful in understanding environmental situations and in telling the approach of danger, and sound-imitation words (sound-related onomatopoeia) are important expressions to inform such sounds in human communication, especially in Japanese language. In this paper, we design a method to recognize sound-imitation words (SIWs) for environmental sounds. Critical issues in recognizing SIW are how to divide an environmental sound into recognition units and how to resolve representation ambiguity of the sounds. To solve these problems, we designed three-stage procedure that transforms environmental sounds into sound-imitation words, and *phoneme group expressions* that can represent ambiguous sounds. The three-stage procedure is as follows: (1) a whole waveform is divided into some chunks, (2) the chunks are transformed into sound-imitation syllables by phoneme recognition, (3) a sound-imitation word is constructed from sound-imitation syllables according to the requirements of the Japanese language. Ambiguity problem is that an environmental sound is often recognized differently by different listeners even under the same situation. Phoneme group expressions are new phonemes for environmental sounds, and they can express multiple sound-imitation words by one word. We designed two sets of phoneme groups: "a set of basic phoneme group" and "a set of articulation-based phoneme group" to absorb the ambiguity. Based on subjective experiments, the set of basic phoneme groups proved more appropriate to represent environmental sounds than the articulation-based one or a set of normal Japaneses phonemes.

1. はじめに

人間は聴覚情報や視覚情報にもとづいて外界の状況を理解しており，特に異常察知の点では聴覚情報は視覚情報よりも重要であると言われている．実際，我々が日常生活において身の回りの音（環境音）を通じて異常に気付いたり，状況を把握したりする機会は非常に多い．例えば，ドアの開閉音や足音を聞いて 誰かが接近していると判断したり，サイレンの音から緊急自動車の通過を知ったりする．このように我々は視覚情報や音声（人間の声）からだけではなく，各種環境音からの情報もまた多く利用している．近年のマンマシンコミュニケーションの多様化に伴い，これらの環境音情報を人間だけではなく計算機もまた理解し利用することが求められるようになった．Jahns らは農場における家畜管理の効率化を目指して牛の鳴き声から個体の認識と各個体の健康状態の把握を自動で行うシステムを開発した [Jahns 98]．また，芦谷らは鳴き声から鳥の種類を認識するシステム [芦谷 92] や環境音を用いて状況を判断する防犯システム [Ashiya 96] を開発している．他にも，音声情報と環境音情報を統合してビデオの自動インデキシングを行う Zhang らの研究 [Zhang 98] なども環境音研究の 1 つとして挙げられる．

本研究では，人間・計算機間における実世界情報の共有化を目指して，環境音を擬音語として自動認識するシステムを設計する．擬音語は日常生活において環境音を扱う際に頻繁に用いられる表現であり，特に日本語は他国語に比べてその頻度が高いと言われている [山口 02]．実

際，和氣らの行った聴取実験結果は，人間が他の人間に音の情報を言葉で伝える時に擬音語を用いた表現を頻繁に用いること，そして，擬音語は他の表現による伝達方法よりも的確に対象音を伝達することを示している [Wake 01].

擬音語認識の最大の利点は，マンマシンインタラクションがより自然で人間らしいものになることである．例えば「今鳴っているトントンという音は何ですか？」のように，計算機が音を擬音語で表現できれば，人間と自然な形で周囲情報のやりとりが可能となる．また，擬音語は，音を聴取者が母国語の音素体系（我々の場合は日本語）にマッピングしたものである．つまり，多種多様な音を，聴取者が認知したとおりに，母国語の音素体系を用いてシンボル化したものと考えることができる．そのため擬音語を中間表現とすることで，計算機は人間の聞こえ方に即して多様な音の間の類似度を計算できるようになる．本稿では，擬音語認識を用いたアプリケーションの例として 4 章で音検索システムについて報告する．

擬音語認識を行う上で音声と環境音の相違点から 2 つの問題が生じる．1 つは，環境音と音声の発音構造の違いから生じるセグメンテーションの問題である．環境音の音響波形は音声のように子音や母音に相当する音響波形が明確には存在しないため，音声と同じ手法をそのまま適用することはできない．もう 1 つの問題は，音声の場合その書き起こしは一意に定まるのに対して，環境音を表す擬音語はしばしば聴取者に依存し，その表記に揺れが生じることである（音素決定曖昧性問題）．例えば，ある聴取者が「バーン」と聞く衝突音を，別の聴取者は「ダーン」「ボーン」「ドーン」などと別の擬音語で表すかもしれない．

本稿では，前者のセグメンテーション問題については，まず音響波形を音節構造に切り分けてから各音節に対して音素認識を行う（2 章）．後者の音素決定曖昧性問題については，複数の日本語音素を一度に表す表現である音声素グループを導入することで擬音語を適切に表現する（3 章）．4 章では擬音語認識を利用して構築した音検索システムを説明する．なお，本研究で扱う音響データはすべて単音とし，混合音から単音への分離問題については本稿では扱わない．また，対象とする擬音語は日本語の擬音語に限定した．

2. セグメンテーション問題

2·1 セグメンテーション問題とは

セグメンテーション問題とは，擬音語認識においてその音素をどのように音響波形から切り出すか，という問題である．現在，多くの音声認識手法は音節構造と音素を HMM により同時に決定している（図 1: 左）．だが，環境音に対して音声認識と同じ手法を用いると図 2 のように必要以上に多くの音素を切り出し，認識結果は人間

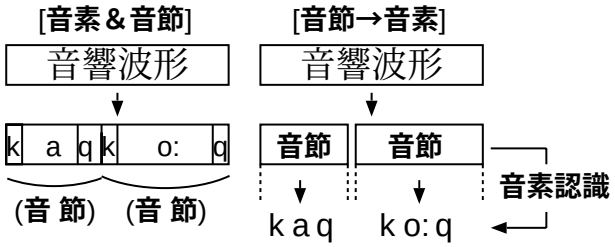

図 1 セグメンテーション問題（左: 通常の音声認識，右: 提案手法)

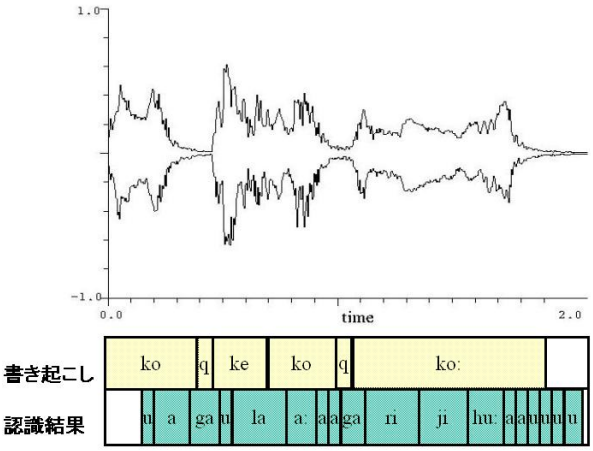

図 2 鶏の鳴き声の波形，人間の書き起こし，自動認識結果

の聞こえ方と大きく異なってしまう．

これは，環境音は音声と発音原理が異なるためである．具体的には，環境音の音響波形の中には音声の音素（特に母音）に相当する部分が含まれていない場合が多いこと，そして，環境音は音声と比べて 1 つの音素を示す音響波形の音長が数十ミリ秒から数十秒まで大きく変動することが理由として挙げられる．擬音語とは聴取者が環境音の音響波形に無理やり日本語の音素を当てはめたものであるということができる．例えば，白色雑音のように音響波形上では子音の連続であるような環境音を聞いた時でも，我々は本来存在しない母音を自動的に補完して表現している．結果として，擬音語を発話した音声の構造は対応する環境音の構造と大きく異なる．

このセグメンテーション問題に対して本研究では，まず音節構造の同定を先に行い，特定した音節範囲ごとに音素を認識する方法を用いる（図 1: 右）．音節構造は音素とは異なり聴取者に依存しない傾向が強い．我々はこのことを聴取実験結果から確認しており [Ishihara 03]，さらにこれは音響波形のパワーと音節の関係を示した Sonority 理論 [Harcourt 93] とも一致する（2·2 節）．聴取者の曖昧性の影響が少ない音節構造を先に決定することで，音素を多く切り出しすぎてしまうセグメンテーションエラーやそれに伴う音素認識ミスを防ぐことができる．

2·2 音節構造に着目した擬音語認識処理

本手法の具体的な処理の流れを以下に示す（図 3）．

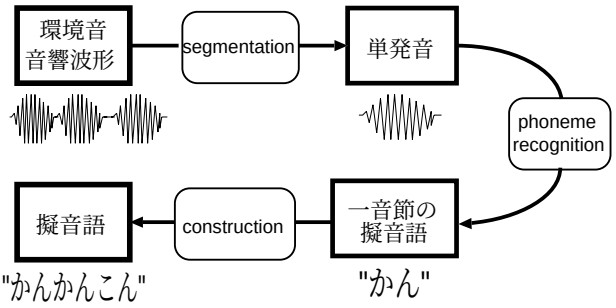

図 3 擬音語認識の処理の流れ

(1) 入力環境音を音節ごとに切り分ける.(segmentation)
(2) 切り分けた単発音を音素認識によりそれぞれ単音節の擬音語に変換する.(phoneme recognition)
(3) 単音節の擬音語を統合する.(construction)

ここで，単発音とは短時間で減衰する音を指す言葉であり [比屋根 98]，本研究においては，擬音語の 1 音節に相当すると見なしている．音節の定義は [子音 + 母音 + 促音 or 撥音] とした．言語学の定義ではこの他に [子音 + 母音] のパターンも存在するが，音節末に促音や撥音がつくか否かは音節後の無音区間長の影響を強く受けるため，各音節ごとに決定することができない．そこで，すべての音節末に促音・撥音が存在するものとして認識し，各音節を統合するステップ 3 において促音・撥音の有無を決定する．以下，各ステップの処理について詳しく説明する．

§1 ステップ 1: Segmentation

ステップ 1 では，擬音語に変換した時に各セグメントが 1 音節の擬音語に相当するように音響波形の切り出しを行う．我々は聴取実験と Sonority 理論 [Harcourt 93] から,「人が聞く擬音語の 1 音節は，音響波形のパワー包絡の 1 山に相当する」という仮説を立てた．この仮説に基づいたセグメンテーション手法を以下に示す（図 4）.

(1) 音響波形のパワー包絡を計算.
(2) 包絡のピークを算出してピーク集合を設計する.
(3) 隣接するピークの小さい方と，ピーク間の谷のパワーの比を計算.
(4) 比の値が閾値以上であれば谷のインデックスで切り出しを行い，以下であれば低い方のピークをピーク集合から外す.
(5) 3 に戻る.

無作為に選出した環境音 [Source 1, Source 2] のセグメント点 157 点に対して本セグメンテーション手法を適用した結果を表 1 に示す．比較として音声 HMM を用いた通常の音声認識システム，および，3,795 サンプルの環境音で作成した HMM でも同様の実験を行った．結果として，音声認識システムや環境音 HMM は人間の聞こえ方よりも 3 倍から 4 倍ほど多くの音節を認識している一方で，本手法は人間の聞こえ方に近い音節切り出しを行っていることが分かる.

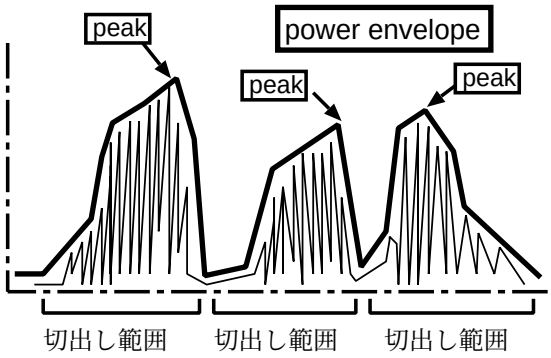

図 4 ステップ 1 : Segmentation

表 1 セグメンテーション比較実験（再現率／適合率）

ピーク単位	音声認識システム	環境音 HMM
83.7%/99.1%	100.0%/26.2%	89.1% / 38.9%

§2 ステップ 2: Phoneme recognition

ステップ 2 では，分割したそれぞれのセグメントに対して音素認識を行い，単音節の擬音語に変換する．音素認識を行う上で問題となる音素決定曖昧性問題については 3 章で詳しく述べる．ここで必要となる音素認識は，音信号を人間の聞こえ方に対応する音素へと変換するという点で，音声認識と同じである．したがって本稿では，音素認識を行う特徴量・手法として，音素を識別して表現する枠組みとして広く用いられている MFCC と HMM を採用した．ステップ 1 でセグメンテーション問題を解決しているため，この段階では音声認識と同様の手法を採用しても問題は生じない．実際，予備実験において幾つかの手法や特徴量と比較を行った結果から，HMM と MFCC は最も高い適合率・再現率を示すことを確認している．なお MFCC は，環境音の音源同定のタスクにおいても現時点で最も有効な特徴量の 1 つであると報告されている [Cowling 03].

§3 ステップ 3: Construction

ステップ 3 では，ステップ 2 で得た単音節の擬音語を再統合して，最終出力である複数音節の擬音語を生成する．統合時に，各音節間の無音区間長に応じて促音や撥音の有無を決定する．その閾値は聴取実験結果から 400ms に定めた．現時点では単純な規則で整合するのみであるが，発展課題として統合により得た擬音語を日本語の慣習・言語に基づいて補整することも考えられる．だが，音声・言語学の分野ではまだこれらの影響に関する知見は充分には得られておらず，また擬音語のコーパスに相当するものが存在しないという理由から，本稿では音響的情報のみにもとづいた「聞こえたまま」の擬音語に変換するまでを対象としている．将来的には「言語モデル」としてこれらの文化的・言語的情報を用いた擬音語認識を実現する予定である.

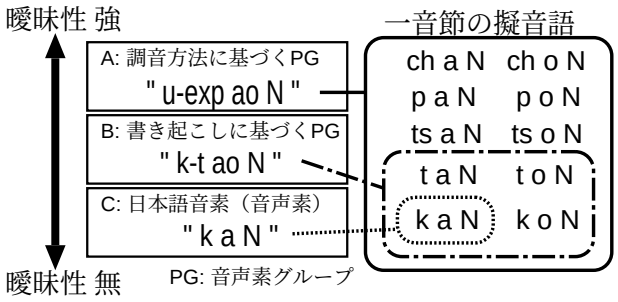

図 5 各音素集合の表現の例と対応する擬音語

3. 音素決定曖昧性問題

3·1 環境音素の設計

環境音の擬音語認識を行う上では，音素決定における曖昧性もまた大きな問題である．環境音を表す擬音語は 1 つではなく，聞く人間や聞く状況に応じて様々な擬音語表現が用いられる．そのため，環境音を表す擬音語が一意に定まらず，認識結果がユーザによって適切であったり不適切であったりするという「聴取者依存性」が生じる．本章ではこの問題を解決するために，複数の擬音語を一語で表す表現を可能にする環境音用の音素を設計する．区別のためこの新しい音素を**環境音素**と呼び，従来の音声言語の日本語音素を**音声素**と呼ぶ．我々は環境音素の候補として以下の 3 種類の音素集合を用意した．

- $\mathcal{A}$: 調音方法に基づく音声素グループの集合
- $\mathcal{B}$: 書き起こしに基づく音声素グループの集合
- $\mathcal{C}$: 日本語音素（音声素）の集合 （表 2）

$\mathcal{A}$ と $\mathcal{B}$ の音素は複数の音声素をグループ化したものであり，**音声素グループ**と呼ぶ．例えば，/b/とも/d/とも聞こえる音を音声素グループ/α/で表し，/a:/とも/o:/とも聞こえる音を音声素グループ/β/で表すとすると，“$\alpha-\beta-N$” という表現は「バーン（ba:N）」「ボーン（bo:N）」「ダーン（da:N）」「ドーン（do:N）」の 4 種類の擬音語を表すことになる．このように複数の擬音語を表す表現を用いることは，複数のユーザの複数の回答を含んだ表現（曖昧性を許容した表現）を実現することである．ただし，曖昧性許容度が大きすぎて，どのユーザの表現とも一致しない擬音語を多数生成する音声素グループは不適切である．そこで，どの音素集合が最もユーザの表現の曖昧性を適切に表し，環境音素として妥当であるかを決定するために各音素集合を比較する必要がある（図 5）．$\mathcal{C}$ は音声認識で用いられる日本語音素の集合（表 2）であり，音声素グループに対する比較として用いる．以下，2 種類の音声素グループの設計方法を説明する．

[音素 $\mathcal{A}$] 調音方法に基づく音声素グループ

$\mathcal{A}$ は，音声発話における子音の調音方法に着目した音声素グループである．我々は音象徴の理論 [田守 99] に基づいて，「人間は環境音を擬音語で表す際に，同じ調音方法で発音される子音を混用しやすい」という仮説を立てた．例えば，「バーン」「ダーン」など，爆発を表す擬音語の子音には有声破裂音を用いることが多い．この仮説に基づき，調音方法による音声素分類と聴取実験結果から表 3 のように子音の音声素グループを設計した．これを**調音方法に基づく音声素グループ（articulation-based phoneme groups : APGs）**と呼ぶ．母音は聴取実験の結果から，/ao/，/i/，/u/，/e/，/ao:/，/i:/，/u:/，/e:/ の 8 クラスとした．なお，/ao/は/a/とも/o/とも聞こえる音を指す．

表 2 日本語音素（音声素） [$\mathcal{C}$]

/a/, /i/, /u/, /e/, /o/, /a:/, /i:/, /e:/, /u:/, /o:/, /N/, /w/, /y/, /p/, /t/, /k/, /b/, /d/, /g/, /ts/, /ch/, /m/, /n/, /h/, /f/, /s/, /sh/, /z/, /j/, /r/, /q/

表 3 調音方法に基づく音声素グループ の子音 [$\mathcal{A}$]

音声素グループ	対応する日本語音素	調音
/nasal/	m n	鼻音
/fric/	j s sh z	摩擦音
/hf/	f h	摩擦音
/semiv/	w y	半母音
/v-exp/	b d g gy	有声破裂音
/u-exp/	ch k p t ts	無声破裂音

この音声素グループの子音は 6 クラスしかなく，個々のクラスは多くの音声素に対応しているため，APG を用いた表現は他の音素を用いた表現よりも擬音語を多く生成する（曖昧性の許容度が大きい）．つまり，出力する擬音語集合は多くのユーザが表す擬音語を含むものの（高再現性），逆に，いずれのユーザにとっても妥当ではない擬音語もまた含みやすい（低適合性）．

[音素 $\mathcal{B}$] 書き起こしに基づく音声素グループ

$\mathcal{B}$ は，音声素のあらゆる組み合わせを認める音声素グループであり，$\mathcal{B}$ の音素集合は，学習用データの書き起こし（人間の付与した擬音語）中に充分な出現頻度（15 回以上）を持つ音声素グループで構成される．例えば書き起こしの中に，/k/としても/t/としても聞こえるが，それ以外の音声素としては聞こえない音が 15 サンプル以上あった場合，/k-t/という音声素グループを音素集合に追加する．これらの音声素グループを**書き起こしに基づく音声素グループ（basic phoneme groups : BPGs）**と呼ぶ．本研究で利用したデータ [Source 3] の書き起こしに現れた BPG を表 4 に示す．BPG は APG とは異なり，グループを構成する各音声素は複数の音声素グ

表 4 書き起こしに基づく音声素グループの音素 [$\mathcal{B}$]

/t/, /k-t/, /b/, /p/, /t-ch/, /sh/, /k/, /f-p/, /t-p/, /z-j/, /g/, /r/, /k-p/, /ch/, /k-t-ch/, /b-d/, /j/, /t-ts/, /w/, /ts-ch/, /s-sh/, /k-t-r/, /d-g/, /b-d-g/, /sh-j/, /k-g/, /t-d/, /ao/, /a/, /i/, /u/, /e/, /o/, /ao:/, /a:/, /i:/, /u:/, /e:/, /o:/, /N/, /Q/, /Q-N/

表 5 音素認識器の設計条件

認識器	HMMs（16 混合モノフォンモデル）
学習データ	6,011 音 [Source 3] サンプリング周波数 44,100Hz 擬音語ラベルは人手で付与
特徴量	MFCC(16)+Pow+ΔMFCC(16)+ΔPow frame size: 75 ms，frame shift: 15 ms
デコーダ	HVite in HTK [HTKBOOK]

ループに重複して属しても良い．

出現頻度（学習サンプル数）が 15 に満たない BPG で表記される音響データは，その BPG の代わりに BPG を構成する音声素の表現を 1 つ 1 つ個別に用いて学習を行う．例えば，/p-w/はデータ中の学習サンプル数が不充分であるため，/p-w/という BPG を子音としたラベルでは学習を行わず，/p/を子音として作成したラベルと/w/を子音として作成したラベルをそれぞれ学習することで対処している．

3·2 評 価 実 験

本節の目的は，3 種類の音素集合の表現を比較評価し，ユーザ間の表現の曖昧性を適切に表すことができる音素集合を,「環境音を擬音語として表すのに最も適切なシンボルの集合（環境音素）」として定義することである．そのために 3 種類の音素集合からそれぞれ音素認識器を作成し，その認識結果を用いて評価実験を行う．認識器の設計条件を表 5 に示す．MFCC の 0 次項を含めているのは，環境音は音声の場合とは異なりパワーが音素決定に影響するためである．

本評価実験では各音素集合の評価基準として，それぞれのシステムが出力した擬音語の『再現率』と『適合率』，そして，その擬音語が対象音を表す擬音語として適切であるかを被験者が評価する『評点』の 3 項目を用意した．すなわち，適切な音素集合とは『様々な被験者が回答した擬音語を多く含み（再現性），同時に，どの被験者の擬音語とも一致しないような擬音語は含まない（適合性）擬音語集合』を生成する音素集合である．再現率と適合率の定義を以下に示す．

表 6 評価実験結果

	再現率		適合率		評点
$\mathcal{A}$	81/140	(57.9%)	27/104	(26.0%)	—
$\mathcal{B}$	79/140	(56.4%)	26/36	(72.2%)	3.89
$\mathcal{C}$	56/140	(40.0%)	17/22	(77.3%)	3.66

$$\text{再現率} = \frac{\text{認識結果内に一致する擬音語が存在する回答数}}{\text{被験者の回答総数}}$$

$$\text{適合率} = \frac{\text{一致する被験者回答が存在する認識結果の数}}{\text{システムの認識結果総数}}$$

これらの定義は出力・正解が複数存在するという点で通常の再現率・適合率とは異なる．なお，聴取実験において被験者が書き起こした擬音語集合を正解の擬音語とした．

評価実験の被験者は日本語を母語とする 20 代前半の男性 7 人である．評価データには実環境で録音した単発音と，効果音 CD（[Source 1, Source 2]）から抽出した単発音の合計 20 サンプルを用いた．実験の具体的な内容を以下に記述する．

(1) 被験者は環境音を聞く．

(2) 被験者は自分の聞こえ方に従って音を擬音語として書き表す．複数の聞こえ方がある場合は複数の擬音語を記述する.（これを評価時の正解ラベルとする）

(3) その環境音に対して各認識器が出力した擬音語をすべて被験者に提示する．

(4) 被験者は，各擬音語がその環境音を表す音として適切か否かを 1 点（不適切）から 5 点（適切）で評価する．

評価実験の結果を表 6 に示す．曖昧性の許容度が強い表現ほど多くの擬音語を出力するため，再現率は低く，一方で適合率は高くなっている．我々が提案した $\mathcal{B}$ の音声素グループ（BPG）の表現は，擬音語を 1 つしか出力しない $\mathcal{C}$ の音声素の表現に近い適合率を持ち，同時に，非常に多様である聴取者の擬音語表現（表 7）の半分以上を再現している．また，BPG の表現の評点平均は 3.89 であり，通常の日本語音素を用いた表現よりも BPG を用いた表現の方が適切であると聴取者は判断している．なお，$\mathcal{A}$ の音声素グループ（APG）を用いた表現は非常に多くの擬音語を生成するため，聴取者に全ての擬音語の評点を行わせることが難しく，今回の実験では評点平均を算出していない．ただ，他の 2 表現と共通する擬音語の評点や再現率・適合率を考慮すると，APG の評点平均は 3.00 以下であると推測する．

これらの実験結果から我々は，実際の人間の書き起こしから生成した BPG の表現は，通常の日本語音素の表現や調音方法という知識に基づいて設計した音素集合の

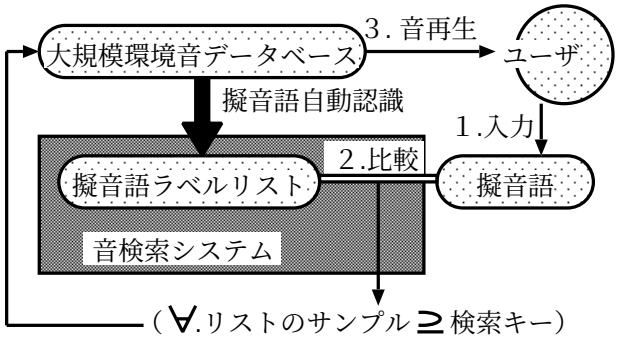

図 6 環境音検索の処理

表現に比べて，人間の表す擬音語の曖昧性を適切に表現しており，環境音素として適当であると判断した．各システムの出力結果と聴取者の表した擬音語の回答を表 7 に示す．上段が各システムの出力であり，下段は聴取者の回答である．丸括弧内の数値は対象の評点平均を，角括弧内の数値はその擬音語を回答した聴取者の人数を表す（複数回答を認めているため合計は被験者数と一致しない）．

この表では，音声素の表現が洩らした聴取者の擬音語の一部を，BPG の表現がカバーしていることが分かる．例えば，No.05 のサンプルは，聴取者によって「きん」「てぃん」「ちん」と様々な擬音語で表されているが，音声素の表現は「ちん」という擬音語しか表していない．このようなラベルを次章で述べる音検索システムのラベルとして適用した場合,「きん」と聞こえたユーザはこの音を検索で取得できないことになる．一方で，BPG の表現は上位 3 つの擬音語をカバーしているため，この表現をラベルとすることで多くのユーザが各自の擬音語を検索キーとして対象音を取得することができる．また，APG の表現では曖昧すぎて，聴取者の回答には無い「ぴん」「つぃん」などの擬音語も生成してしまう．そのため，APG の表現を検索ラベルとして用いた場合，検索結果には不適切な候補が大量に出力されてしまう．全実験結果と音響データは，http://winnie.kuis.kyoto-u.ac.jp/members/ishihara/onomatopoeia.html で公開している．

4. 擬音語認識に基づいた音検索システム

擬音語認識の一応用例として，擬音語認識によりラベルを自動付与した環境音検索システムについて報告する（図 7）．擬音語を用いた環境音検索システムとは

- 「『キーンカンカンカン』と鳴る音が欲しい」
- 「『ザーッ』という音が 3 回以上続く音が欲しい」
- 「『チッ』という音を含むが『トン』という音を含まないような音データが欲しい」

のようなユーザの擬音語を用いた要求に対して，合致する環境音の音情報を出力するシステムである．

表 7 評価実験のデータおよび実験結果

No.	$\mathcal{A}$ の表現	$\mathcal{B}$ の表現	$\mathcal{C}$ の表現
	被験者回答		
01	u-exp ao Q (-)	t ao N(3.45)	t ao Q(3.22)
	こっ[4] かっ[3], とっ[2], とん [2], とぅん [1], こん [1], たっ[1]		
02	u-exp i N (-)	t-ch i N(3.45)	ch i N(3.22)
	てぃん [3], きん [3], ちん [3], かん [1], とぅいーん [1]		
03	u-exp ao N (-)	k-t ao N(3.39)	t ao N(3.06)
	かーん [4], たん [2], かん [1], とーん [1] こーん [1], ぱーん [1], たーん [1], くぁーん [1]		
04	u-exp ao q (-)	k-t ao q (2.97)	t o q(2.00)
	かっ[4], たっ[4], ちゃっ[1], たん [1], ちっ[1], てっ[1], つぁっ[1]		
05	u-exp i N (-)	k-t-ch i N (4.33)	ch i N(4.67)
	きん [4], てぃん [3], ちん [3], ちっ[2], てぃっ[2], とぅっ[1], ぴっ[1]		
06	u-exp i: N (-)	t-ch i: N (4.45)	ch i N (3.67)
	ちーん [4], ちん [3], てぃーん [3], きーん [3], きん [2], ぴん [1]		
07	fric u: q (-)	s-sh u: q (3.50)	s u: q (2.89)
	しゅーっ[5], しーっ[5], しゃーっ[1], じーっ[1]		
08	u-exp o q (-)	p ao q (3.45)	p o q (3.33)
	ぱっ[6], ぽっ[4], たっ[1], かっ[1], ぽわっ[1], くっ[1], つっ[1], とぅっ[1]		
09	u-exp ao q (-)	t ao q (3.06)	t o q (2.56)
	たっ[3], とっ[3], ぱっ[2], かっ[2], ぷっ[1], とぅっ[1], かん [1], こん [1], ぴっ[1]		
10	u-exp i: q (-)	f-p i: q (3.78)	f i: q (3.00)
	ぴーっ[5], ぴぃーっ[2], きゅぃーっ[1], きーっ[1], ぴゅぃーっ[1], ちーっ[1], ふぃーっ[1]		

現状の多くの音検索システム*1は「音源名」や「音長」でしか検索できず，探したい音の正確な音源名が分からない場合には利用できない． このような問題から，和氣らは「音源名」の他に「擬音語」と「形容詞」を検索フィールドとして利用することを提案している [Wake 01]．だが和氣らのシステムは，擬音語や形容詞の曖昧性については考慮しておらず，また，各サンプルのラベルをすべて人手で付与しなければならない．本稿で提案した擬音語認識手法はこれらの『ラベル自動付与』と『ラベルの曖昧性』の両方の問題を解消しており，大規模データベースに対する音検索システムを設計する上で非常に効果的である．

ここでは，RWCP 環境音データベースの単発音（1,898 サンプル）[Source 3] とそれらの単発音を合成した複合音（3,786 サンプル）を検索対象とするシステムを構築した．検索対象に対しては擬音語自動認識を行い，得た

*1 例えば，Sound Effect Searcher（http://www.speed.co.jp/hpa/se/index.htm）などがある

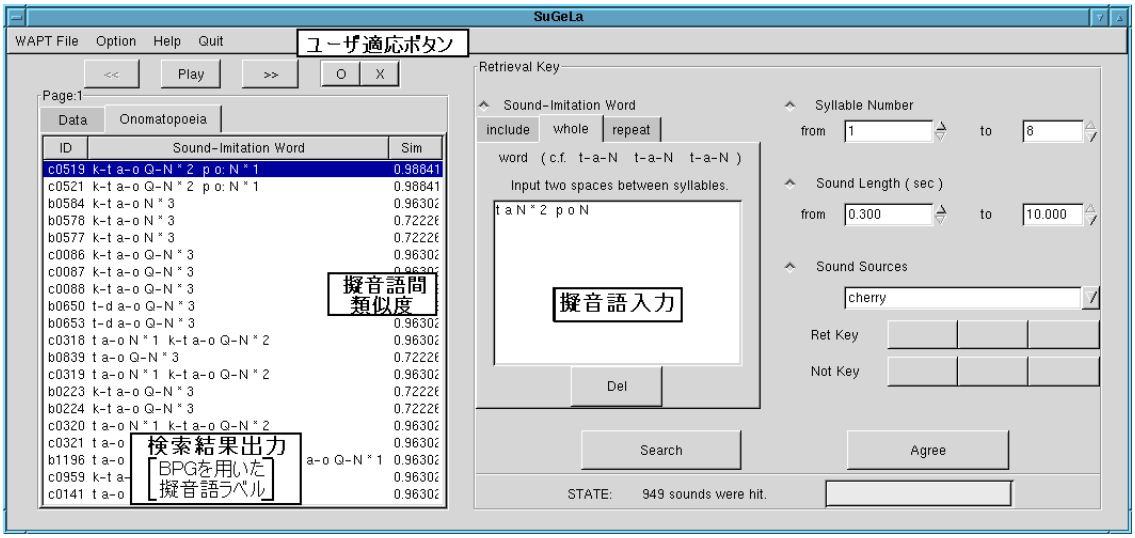

図 7 環境音検索システム

擬音語をあらかじめラベルとして付与しておく．システムは，ユーザから入力された擬音語と付与されている擬音語ラベルを比較することにより，ユーザの要求に近い音を出力できる（図 6）．なお，本システムは「擬音語」だけではなく「音源名」「音節数」「音長」などの要素を加えて検索することもできる．

類似擬音語検索

擬音語表現の曖昧性は環境音素によりある程度解消したが，適合率の問題からすべてのユーザの表現をカバーしているわけではない．そのため一部のユーザの擬音語（検索キー）では，そのユーザが求める環境音を検出できないことがある．この問題を解決するために本検索システムは，検索キーと完全一致しないラベルに対しても擬音語間類似度を計算して，充分に高い類似度を持つサンプルについてはユーザの要請に応じて出力する『類似擬音語検索機能』を実装している．類似度は音素の種類（子音，母音，促音・撥音）の重要度と音素間類似度から計算する．音素間類似度は音声素グループを利用して初期値を決定している．これらの定数や音素間類似度は，ユーザの反応（評価）に応じて値を変動させることもできる（ユーザ適応機能）．それらの詳細については別途報告する．

5. ま と め

本稿は人間・計算機間の実環境情報の共有化を目指して，環境音の擬音語自動認識手法を設計した．本研究の意義を以下にまとめる．

- これまで「音源名」というシンボルを介してのみ扱われていた環境音に対して，人間が日常生活で用いている「擬音語表現」というシンボルを新たに導入し，より自然で人間らしいマンマシンインタフェースを実現した．これにより正確な「音源名」が分からない時であっても，擬音語を介して音情報を計算機とやり取りすることができるようになり，人間・計算機間のコミュニケーションの幅が大きく広がった．本稿では応用例として，擬音語認識でラベルを自動付与した環境音検索システムを報告している．
- 環境音から言語的情報を抽出する上での問題を，音響的側面（セグメンテーション問題）と言語的側面（音素決定曖昧性問題）の両面から明確化し，さらに，それぞれの問題を解消するアプローチを提案した．擬音語と音響的特徴を論じた発表はこれまでにも数件存在するが，実際にこれらの問題を明確化して認識器を作成したのは本研究が初めてである．セグメンテーション問題に対しては，まず音節構造の認識を行って特徴量の抽出範囲を特定してから音素認識を行う手法を用いることで対処した．曖昧性問題に対しては，複数の音声素を一度に表現できる音声素グループを利用することで問題を解決した．

今後の発展課題として，擬音語に対する文化・言語的影響の考慮，混合音から自動で認識対象とする音を抽出する手法について現在検討している．

謝 辞

本研究は科学研究費補助金，NICT，および 21 世紀 COE プログラムの支援を受けた．また，RWCP の実環境音声・音響データベースの非音声音ドライソースを利用した．研究にあたって貴重な助言をくださった NTT CS 研の中谷智宏氏，京都大学奥乃研究室の北原鉄朗氏，京都大学学術メディアセンターの坪田康氏に深く感謝する．

◇ 参 考 文 献 ◇

[芦谷 92] 芦谷武彦, 中川正雄: 鳴き声による鳥の種類の認識システム, 電子情報通信学会技術研究報告 SP92-13, 1992.

[Ashiya 96] Takehiko Ashiya, Masafumi Hasegawa, Masao Nakagawa: IOSES: An Indoor Obsevation System Based on Environmental Sounds Recognition Using a Neural Network, *Trans. of the Institute of Electrical Engineers of Japan*, Vol.116-C, No.3, pp.341-349, 1996.

[Cowling 03] Michael Cowling and Renate Sitte: Comparison of techniques for environmental sound recognition, Pattern Recognition Letters 24 pp.2895-2907, 2003
[Harcourt 93] P. Ladefoged: *A Course In Phonetics*, Harcourt Brace College Publishers, 1993.
[比屋根 98] 比屋根一雄, : 単発音のスペクトル構造とその擬音語表現に関する検討, 電子情報通信学会技術研究報告, SP97-125, 1998.
[HTKBOOK] HTK3.0: http://htk.eng.cam.ac.uk/
[Ishihara 03] Kazushi Ishihara, Yasushi Tsubota, and Hiroshi G. Okuno: Automatic Transformation of Environmental Sounds into Sound-Imitation Words Based on Japanese Syllable Structure, *Proc. of EUROSPEECH-2003*, pp.3185-3188, 2003.
[Jahns 98] Gehard Jahns, Wojciech Kowalczyk and Klaus Walter: Sound Analysis to Recognize Individuals and Animal Conditions, *XIII CIGR Congress on Agricultural*, 1998.
[田守 99] 田守 育啓: 『オノマトペ – 形態と意味 –』, くろしお出版, 1999.
[田中 95] 田中基八郎: 異音の表現における擬音語の検討, 日本機械学会論文集 C 編, Vol.61, No.592, 1995.
[Wake 01] Sanae Wake and Toshiyuki Asahi: Sound Retrieval with Intuitive Verbal Descriptions, IEICE 2001, *Tran. on Information and Systems*, Vol.E84-D, No.11, pp.1568-1576, 2001.
[山口 02] 山口仲美 著『犬は「びよ」と鳴いていた - 日本語は擬音語・擬態語が面白い-』 光文社新書, 2002.
[Zhang 98] Tong Zhang and C.C. Jay Kuo: Audio-guided audiovisual data segmentation, indexing, and retrieval, *Proc. of the SPIE*, The International Sociery for Optical Engineering, 3656, pp.316-327, 1998.
[Source 1] 効果音大全集, KING RECORD.
[Source 2] 新・効果音大全集, KING RECORD.
[Source 3] RWCP 実環境音声・音響データベース, http://tosa.mri.co.jp/sounddb/index.htm

〔担当委員：溝口 博〕

2004 年 10 月 29 日 受理

著 者 紹 介

石原 一志(学生会員)

2003 年京都大学工学部情報学科卒業．現在，京都大学情報学研究科修士課程在学．情報処理学会学生会員．

駒谷 和範(正会員)

1998 年 京都大学工学部情報工学科卒業．2000 年 同大学院情報学研究科知能情報学専攻修士課程修了．2002 年同大学院博士後期課程修了．同年より 京都大学情報学研究科助手．京都大学博士（情報学）．情報処理学会平成 16 年度山下記念研究賞受賞．FIT2002 ヤングリサーチャー賞受賞．電子情報通信学会，情報処理学会，言語処理学会各会員．

尾形 哲也(正会員)

1993 年早稲田大学大学院理工学部機械工学科卒業．1997 年日本学術振興会特別研究員，1999 年早稲田大学理工学部助手，2001 年理化学研究所脳科学総合研究センター研究員を経て，2003 年より京都大学大学院情報学研究科知能情報学専攻講師，現在に至る．博士（工学）．2001 年より早稲田大学ヒューマノイド研究所客員講師．人間ロボット音声協調，マルチモーダル知覚、音響信号によるロボット動作生成などの研究に従事．情報処理学会，日本ロボット学会，日本機械学会，IEEE などの会員．

奥乃 博(正会員)

1950 年生 (HAL と同じ誕生日)．1972 年東京大学教養学部基礎科学科卒業．日本電信電話公社，NTT，科学技術振興事業団，東京理科大学理工学部情報科学科を経て，2001 年 4 月より京都大学大学院情報学研究科知能情報学専攻 教授．博士 (工学)．この間，スタンフォード大学客員研究員，東京大学工学部客員助教授．人工知能，音環境理解，ロボット聴覚の研究に従事．1990 年度本学会論文賞，IEA/AIE-2001 最優秀論文賞，IEEE/RSJ IROS-2001 Best Paper Nomination Finalist，第 2 回船井情報科学振興賞等受賞．情報処理学会，日本ソフトウェア科学会，日本認知科学会，日本ロボット学会，ACM，AAAI，IEEE，ASA 各会員．本学会理事．著編書:『インターネット活用術』(岩波書店)，『*Computational Auditory Scene Analysis*』(共編，LEA)，『*Advanced Lisp Technology*』(共編，Taylor & Francis) 他．