1E2-03

環境音の擬音語変換のための環境音用音素の設計

Designing Environmental Phoneme for Automatic Sound-Imitation Word Recognition

石原 一志*1 Kazushi Ishihara　中谷 智広*2 Tomohiro Nakatani　駒谷 和範*1 Kazunori Komatani　尾形 哲也*1 Tetsuya Ogata　奥乃 博*1 Hiroshi G. Okuno

*1京都大学 情報学研究科 Graduate School of Informatics, Kyoto University

*2NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories

Sound-imitation words (SIWs), or onomatopoeia are important in informing sound events in human-computer communication. We designed three-stage procedure for environmental sound recognition: (1) a waveform is divided into some chunks, (2) the chunks are transformed into sound-imitation syllables by phoneme recognition, (3) a sound-imitation words is costructed from sound-imitation syllables according to the requirements of the Japanese language and culture. The main problems in this automatic SIW recognition are ambiguities in determining phonemes, since an environmental sound is often recognized differently by different listeners even under the same situation. To solve this problem, we designed a set of new phonemes, called "basic phoneme-group" set, to represent environmental sounds in addition to an existing set of the "articulation-based phoneme-groups." Based on the subjective experiments, the set of basic phoneme-groups proved more adequate to represent environmental sounds.

1. はじめに

近年の計算機技術の発展は実社会における計算機の応用の場の多様化を促し, それにより音研究は, 音声（人間の声）だけではなく日常における身の回りの音（環境音）もまた重要な研究対象として捉えるようになった. 実際, 環境音から情報を抽出する研究は近年少しずつ増えている. 例えば, Jahns らは牛の鳴き声から健康状態を測定するシステムを [1], 芦谷らは鳴き声から鳥の種類を認識するシステム [2] や環境音を用いて状況を判断する防犯システム [3] 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 [4] なども環境音研究の一つである.

本研究は, マンマシンインタラクションにおける環境音情報の導入を目指して, 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. 一般に日本語は, 日常会話において擬音語を用いる頻度が高い言語であると言われており, ロボット対話に擬音語を導入することでコミュニケーションの高度化が期待できる. また, 擬音語認識はシンボルグラウンディング問題における環境音のシグナル・シンボル変換として位置づけることができる. 擬音語を用いた研究としては, 和氣らは環境音アーカイブの検索キーとして擬音語を利用したシステムを開発しており [5], 田中らは異常音を擬音語で表して機械の故障を検出する手法を提案している [6]. 本研究が目指す擬音語認識はこれらの擬音語タグの自動生成としても利用できる.

環境音の擬音語認識を行う際に最大の問題となるのは, 聴取者に依存した擬音語表記の揺れ（曖昧性）である. 例えば, ある人が「バーン」と聴く衝突音を, 別の人は「ダーン」「ボーン」「ドーン」などと表すかもしれない. 本稿では環境音用の音素として環境音素を導入することでこの問題を解決する.

まず 2 章では, 擬音語認識の処理の概要を述べる. 3 章で曖昧性の問題と環境音素を用いた解決方法を示し, 4 章で本手法の評価実験を行う. 最後に 5 章で結論を述べる.

連絡先: 石原 一志, 京都大学 情報学研究科 知能情報学専攻, 〒 606-8501 京都市左京区吉田本町, e-mail: ishihara@kuis.kyoto-u.ac.jp

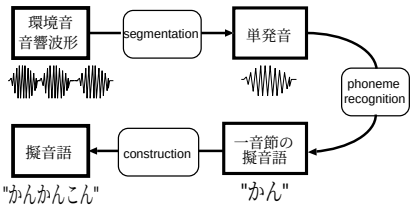

図 1: 擬音語認識の処理の流れ

2. 環境音の擬音語認識

環境音のパワー包絡と擬音語の音節の関係に着目し, 環境音を 3 段階で擬音語に変換する. その具体的な処理の流れを以下に示す（図 1）.

1. 入力環境音を単発音ごとに切り分ける.（segmentation）
2. 各単発音を音素認識により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し, 日本語の慣習に基づいて補整を行う.（construction）

本稿では主に, ステップ 2 の音素認識に関する問題を論ずる. なお, 単発音とは短時間で減衰する音を指す言葉であり [9], 本研究においては, 2.1 節に示すようにパワー包絡の 1 山と同一であると見なしている.

2.1 ステップ 1: Segmentation

ステップ 1 では, 環境音の音響波形をパワー包絡の 1 山ごとに分割する. この分割で得たセグメントはいわゆる単発音と同等であり, 擬音語の 1 音節分に相当することが予備実験と Sonority Theory から判明している [7, 10]. この知見を利用して設計したセグメンテーション手法を以下に示す (図 2)[7, 8].

1. 音響波形のパワー包絡を計算.

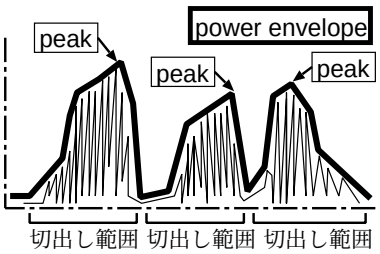

図 2: ステップ 1： Segmentation

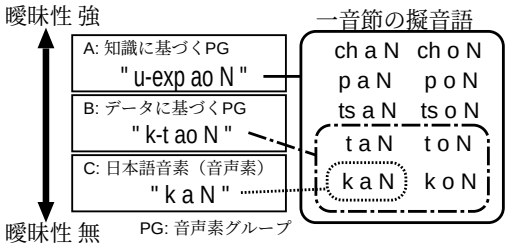

図 3: 各音素集合の表現の例と対応する擬音語

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.

2.2 ステップ 2: Phoneme recognition

ステップ 2 では, 分割した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. システムは 6,011 サンプルの環境音から構築した HMM ベースの認識器で, 16 混合のモノフォンモデルである. 音響データは 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 [S1]. 音響特徴量は 17 次の MFCC（0 次含む）と ΔMFCC の計 34 次元であり, フレーム長 50ms, フレーム間隔 10ms で抽出した. これらの実験条件の値は, 396 種類の条件で行った予備実験で最も高い精度を示した値である. なお, 音素認識を行う際に生じる音素決定の曖昧性問題については 3, 4 章で論じる.

2.3 ステップ 3: Construction

ステップ 3 では, ステップ 2 で生成した単音節の擬音語を統合する. さらにその際に, 得られた擬音語を日本語の慣習・言語に基づいて補整する必要がある. 例えば, 音素認識した擬音語が「かっくー」であったとしても, その音がカッコウの鳴き声であると判明しているのであれば,「かっこー」に修正して出力するのが望ましい. これは,「かっこー」という擬音語がカッコウの鳴き声を表す単語として日本語に浸透しているからである. このような擬音語の決定に影響する文化的・言語的要素を補整戦略として組み込むことについては現在検討中であり, 別途報告する.

3. 音素決定曖昧性問題

環境音の擬音語認識を行う上で最大の問題となるのは, 音素決定における曖昧性問題である. 環境音を表す擬音語は 1 つではなく, 聴く人間や聴く状況に応じて様々な擬音語として表される. つまり, 環境音を表す擬音語が一意に定まらず出力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に対応していく必要がある. 本稿ではこの問題を解決するために, 複数の擬音語を表す表現が可能な環境音用音素を設計する. 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.

表 1: 日本語音素（音声素） [$\mathcal{C}$]

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

3.1 環境音素の設計

環境音素の候補として次に示す 3 種類の音素集合を比較検討する.

- $\mathcal{A}$: 調音方法に基づく音声素グループの集合 （3.1 節）
- $\mathcal{B}$: ケースに基づく音声素グループの集合 （3.2 節）
- $\mathcal{C}$: 日本語音素 (音声素) の集合 （表 1）

$\mathcal{A}$ 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/とも聴こえる音は音声素グループ/α/で表し, /a:/とも/o:/とも聴こえる音は音声素グループ/β/で表すとすると, "$\alpha-\beta-N$" という表現は「バーン」「ボーン」「ダーン」「ドーン」の 4 種類の擬音語を表すことができる. このように複数の擬音語を表す表現を用いることで, 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）を実現することができ, その後の後処理によってユーザごとに適切な擬音語を選択して出力することが可能となる. なお, $\mathcal{C}$ は音声認識で用いている日本語音素の集合（表 1）であり, 実験の比較対象である. 以下, 2 種類の音声素グループの設計方法を説明する.

3.2 音素 $\mathcal{A}$: 調音方法にもと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. 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ことが聴取実験から分かっている [7]. 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が用いられることが多い. この知見に基づき, 調音方法による音声素分類と聴取実験結果から表 2 のように音声素グループを設計した. 母音は聴取実験結果から, /ao/, /i/, /u/, /e/, /ao:/, /i:/, /u:/, /e:/の 8 クラスとした. なお, /ao/は/a/とも/o/とも聴こえる音を指す.

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, $\mathcal{A}$ の表現は他の音素を用いた表現よりも多くの擬音語を生成する（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, いずれのユーザにも妥当ではない擬音語も含みやすい（低適合性）.

3.3 音素 $\mathcal{B}$: ケースにもとづく音声素グループ

$\mathcal{B}$ の音声素グループは音声素の全ての組合わせに基づいて設計した. 例えば/k-t/という音声素グループは, /k/としても/t/

表 2: 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

表 3: ケース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/

としても聴こえるが, それ以外の音声素としては聴こえない音を表す. $\mathcal{B}$ の音声素グループは $\mathcal{A}$ とは異なり, 各音声素が複数の音声素グループに重複して属しても良い. 表 3 に本実験のデータ（6,011 サンプル）中に現れた音声素グループを示す. これ以外のグループは今回の学習データに現れなかったか, 学習を行うためのサンプル数が不足していたか（15 サンプル以下）のいずれかである.

4. 認識システムの作成と評価実験

本稿の目的は, 各音素集合を用いた表現を比較評価し, 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素として定義することである. 一つしか擬音語を表すことができない音声素による表現であっても, スコア閾値を設定することで複数の擬音語を生成することができるが, これは「事象を一意のシンボルに結びつける」というシンボルグラウンディングの観点から適切とは言えない. コミュニケーションで用いるシンボルを定義するという立場から, 本研究では複数のシンボルを出力することを考えず, 単一の出力として最も適切なシンボルの集合（環境音素）を定義する.

環境音素を用いて音素認識した結果から音素集合の妥当性を評価するためには, (1) システムの学習における曖昧性の解消, (2) 音素集合の評価尺度の設計, の 2 点を行う必要がある. 4.1 節, 4.2 節でこれらの問題について論じ, 4.3 節で聴取者を用いた評価実験について述べる. なお, 認識システムの詳細は 2.2 節で説明している.

4.1 重複学習による日本語音素の表現の学習

$\mathcal{C}$:音声素の表現を学習する際に生じる曖昧性問題の解消方法について論じる. 実験の評価対象は各音素集合の表現であり, その評価はシステムの出力結果の曖昧性に基づいて行う. そのためシステム間の性能差が出力結果に影響すると, 適切な評価を行うことができない. 一方, 音声素の表現は音響データを表す複数の擬音語のうち 1 つしか表すことができないため, 同じ音響データに対して複数の擬音語全てを学習できる他の音素集合のシステムとは差が生じる.

我々はこの問題を解決するために, システム設計において重複学習を採用した. 重複学習とは, 対応するラベルが複数存在する場合に, 対象となる音響データをそれぞれのラベルで 1 つずつ複数回学習するという手法である. 重複学習を行うことで $\mathcal{C}$:音声素の表現であっても全ての聴こえ方に対応した学習を行うことができる.

表 4: 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

4.2 評価尺度の設計

コミュニケーションでの利用という観点からすると, 適切な音素集合のシステムとは『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』を出力するシステムである. そこで, 各音素集合の評価は, 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』, 擬音語に対する被験者の『評点』の 3 項目に基づいて行う. 再現率と適合率は下式のように定式化した. 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.

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

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

4.3 評価実験

3 種類の音素集合の評価を目的として 7 人の被験者に対して聴取実験を行った. 評価データには, 実環境で録音した単発音と効果音 CD（[S2, S3]）から抽出した単発音の合計 20 サンプルを用いた. 以下に実験の内容を記述する.

1. 被験者は環境音を聴く.
2. 被験者は自分の聴こえ方に従って音を擬音語として書き起こす. 複数の聴こえ方がある場合は複数の擬音語を書く.
3. その環境音に対して各システムが出力した擬音語がすべて被験者に提示される.
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 4 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の精度を持ち, 一方で, 非常に多様である聴取者の擬音語表現（表 5）の半分以上を再現している. また, $\mathcal{B}$ 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 $\mathcal{B}$ の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

表 5: 聴取実験・擬音語例

No.	$\mathcal{A}$ の表現	$\mathcal{B}$ の表現	$\mathcal{C}$ の表現
	被験者回答		
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		
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		
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		
09	u-exp o q (—)	p ao q (3.45)	p o q (3.33)
	ぽっ6, ぽっ4, たっ1, かっ1, ぽわっ1, くっ1, つっ1, とぅっ1		
10	u-exp ao q (—)	t ao q (3.06)	t o q (2.56)
	ぽっ3, たっ3, とっ2, かっ2, ぷっ1, とぅっ1, かん 1, こん 1, ぴっ1		
11	u-exp i: N (—)	r i N (2.22)	r i: N (4.44)
	りーん 5, ちーん 3, てぃーん 3, ぴーん 3, きーん 2		
13	u-exp o: N (—)	p o: N (4.33)	p o: N (4.33)
	ぽーん 7, こーん 2, べーん 1, ぱーん 1, くぁーん 1, くぉーん 1		
15	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		
17	v-exp u: q (—)	b u: q (4.56)	b u: q (4.56)
	ぶーっ5, びーっ4, ぶぁーっ1, どぅーっ1,		
18	v-exp u: q (—)	g u: q (2.67)	g u: q (2.67)
	じーっ3, ぎーっ2, うぃーっ1, じわじわじわ 1, じぃーっ1, びーっ1, でぃでぃでぃ1		
19	u-exp i: N (—)	t-ch u: q (1.56)	t i: N (4.00)
	ちん 4, てぃーん 3, きん 2, てぃん 2, きぃん 1, きぃーん 1, ちぃーん 1		
20	v-exp u: N(—)	b u: N(4.56)	g u: N (3.44)
	ぶーっ4, ぶーん 2, ぐーん 2, びーっ1, どぅーん 1, ばーん 1, ぼーん 1, ぶわーっ1		

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, $\mathcal{A}$ の音声素グループの評点平均は 3.00 以下であると推測している.

これらの実験結果から我々は, 音声素の全ての組み合わせをカバーした $\mathcal{B}$ の音声素グループの表現は, 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表現に比べて, 人間の表す擬音語の曖昧性に非常に近く, 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例を表 5 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. 括弧内の数値は対象の評点平均を, 回答の後ろの数値はその擬音語を回答した聴取者の人数を表す. 全実験結果および音響データは, http://winnie.kuis.kyoto-u.ac.jp/members/ishihara/onomatopoeia.html で公開している.

5. 結論

本稿では, 3 段階で環境音を擬音語として認識する手法について述べた. 特に, ステップ 2 の音素認識における音素決定の曖昧性問題に焦点を絞り, その解決方法として音声素グループという音素の集合を設計して, 通常の日本語音素と比較評価を行った. 被験者を用いた評価実験により各音素集合の表現の比較評価を行ったところ, ケースに基づいて設計した音声素グループの表現は日本語音素の表現に比べて人間の表す擬音語表現の曖昧性を適切に表しており, 環境音素として妥当であることが分かった. 現在, 音響信号が与えられると, それを模倣する擬音語（SIW）が自動的に得られるシステムが構築されている. 今後の課題としては, ステップ 3 である 1 音節の擬音語の統合・補整において, 聞き倣しなどの文化・言語依存の要素を組み込む手法の検討などが挙げられる. 本研究は科学研究費補助金, 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.

参考文献

[1] Gehard Jahns, Wojciech Kowalczyk and Klaus Walter: Sound Analysis to Recognize Individuals and Animal Conditions, *XIII CIGR Congress on Agricultural*, 1998.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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

[8] 服部 佑哉, 石原一志, 尾形哲也, 奥乃博: 連続環境音の繰返し構造の認識, 情報処理学会第 66 回全国大会, 2004

[9] 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.

[10] P. Ladefoged: *A Course In Phonetics*, Harcourt Brace College Publishers, 1993.

[11] 田守 育啓: 『オノマトペ – 形態と意味 –』, くろしお出版, 1999.

[12] HTK3.0: http://htk.eng.cam.ac.uk/

[S1] RWCP 実環境音声・音響データベース, http://tosa.mri.co.jp/sounddb/index.htm

[S2] 効果音大全集, KING RECORD.

[S3] 新・効果音大全集, KING RECORD.