1C2-OS4b-4

擬音語と環境音の音響的関係性を考慮した環境音-擬音語変換システム

Environmental-Sounds-to-Sound-Imitation-Words Conversion System Considering Acoustical Relations between Environmental Sounds and Sound-Imitation Words

山川暢英[*1] Nobuhide Yamakawa　北原鉄朗[*2] Tetsuro Kitahra　高橋 徹[*1] Toru Takahashi　尾形哲也[*1] Tetsuya Ogata　奥乃 博[*1] Hiroshi G. Okuno

[*1]京都大学大学院 情報学研究科 知能情報学専攻 Graduate School of Informatics, Kyoto University

[*2]日本大学 文理学部 College of Humanities and Sciences,Nihon Univeristy

Recognition of sound-imitation words (SIWs), or onomatopoeic representations, by computers enables us to share the information of "how one hears sounds" among them, and can be a powerful tool for human-robot interactions. In the conventional study, SIWs were computationally recognized by using the common human-speech-recognition features and classifier, despite the inconsistency in psychoacoustical properties. Taking the relation between human sound perception and SIWs into account, we propose a SIW recognition system which uses the temporal and spectral features designed for SIW vowels and consonants individually in accordance with their psychoacoustical properties. Experimental evaluation indicates that low-level spectral features, such as spectral centroid, are more effective for SIW vowel recognition in comparison with MFCC used in the conventional method, whereas errors in the estimations of syllable structures increase due to the inappropriate syllable merging scheme which cannnot handle the period of syllable chunks.

1. はじめに

昨今，ロボット制御技術の進歩に呼応する形で，視聴覚を中心に人工知能の認知機構研究が盛んになっており，実世界における応用の多様化が進んでいる．特に聴覚はこれまで人間の音声認識を軸に発展を遂げてきたが，近年は音声だけでなく日常における身の回りの音（環境音）も認識対象として捉え，そこから有意な情報を抽出し識別及び対話システムへの応用を目指す研究が少しずつ増えている．現在の環境音認識の主流となる研究には，防犯システムへの応用を見据えた異音検知がある [Cowling 03, Ntalampiras 10, Chan 10]．また，雑踏，騒音といった背景音を認識し，自分がどういった環境あるいは場所に置かれているかをコンピュータに理解させ，挙動制御に役立てようとする研究も行われている [Eronen 06, Chu 09]．このように環境音認識研究はそのほとんどが音源同定問題か外れ値検出問題として扱われており，コンピュータに「どの音 (what) が鳴っているか」という情報を抽出させることを目的としている．しかしこの方法では，音が「どのように（how）鳴っているか」といった人間が知覚する音の状態までは表現困難である．

本研究では, 環境音の状態情報 (how) を扱えるマンマシンインタラクションの実現を目指して，環境音の音響信号を擬音語で認識するシステムを設計する．擬音語は人間が聴取した音の迫力，明るさ，荒さといった "音の感覚" を言語でもって抽象化したシンボルであり，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可能である．特に日本語は他言語に比べ日常会話での擬音語使用頻度が高いとされており [筧 93]，計算機が擬音語という人間特有の "音の感覚" を理解できることで，ユーザにとってより対人間に近い自然なコミュニケーションの実現が期待できる．

過去の擬音語の応用事例としては, 和氣らの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 [Wake 01], 田中らの異常音を擬音語で表して機械の故障を検出する手法 [田中 95]，製品が発するビープ音の擬音語表現とその機能イメージとの対応関係を調べ工業デザインに役立てようとする研究 [山内 03]，などが挙げられる．これらは全て限られた音源のみを対象としていたが，石原らにより，特定のドメインに限定されない擬音語を自動出力するシステムが開発された [石原 05]．そこでは，擬音語の特色の一つである「表現の聴取者依存性」（ある環境音の擬音語表現は一意に定まらず聴取者によって変化しうる）に対応する為に，例えば，"コン" と "トン" どちらでも表現されうるような音の子音表現には "$/k-t/$" という音素を用いるなど，同時に複数通りの擬音表現が可能な環境音用の音素グループ（環境音素）を新しく設計した．また，音素グループの設計と音節区間の切り出し手法以外には，識別器に隠れマルコフモデルを使用し，子音・母音・促音・撥音の区別に関係なく，全ての音素に対してメル周波数ケプストラム係数 (MFCC) を採用していた．ほぼ 6 割の確率でユーザにとって正しい擬音語を出力するが，母音の誤認識や扱える擬音語の音節構造が限定的といった問題があった．

詳細は後述するが，人間が擬音語認知をする時，母音の決定は信号のエネルギー分布（e.g., どの帯域に高いエネルギーを持つか）に依存して行われ，濁音の有無などの子音決定は倍音構造などのスペクトル形状（e.g., 音色）と強い相関があることが先行研究により指摘されている [比屋根 98, Takada 06, 大石 09]．この特性の違いから，子音と母音に同じ特徴量を適用する手法が妥当かどうか疑問が生じる．また，HMM などのモデルベースの識別器による擬音語認識は，扱える音節構造が事前にモデル内に埋め込まれるため，モデルの設計段階で扱える擬音語表現の範囲が限定されてしまう．したがって，認識に用いる音響特徴や識別手法も擬音語認識用に設計／選択する必要があり，その際，人間の環境音知覚と擬音語認知の関連性を考慮していることが望ましい．本稿では．擬音語認知に関連した音響心理の知見を援用し，音素種類（母音，子音）と音節構造でそれぞれの特性に合わせた音響特徴を使用して擬音語認識を行い，従来法と認識性能を比較する．

本稿の構成を以下に示す．第 2 章では擬音語認識システムの

連絡先: 山川暢英，京都大学大学院情報学研究科，〒606-8501 京都府京都市左京区吉田本町工学部 10 号館, 0757534952, 0757535977, nyamakaw@kuis.kyoto-u.ac.jp

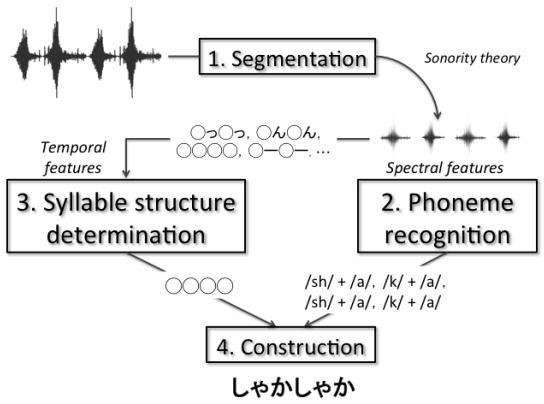

図 1　擬音語認識処理の流れ

概要を示し，第 3 章では特徴量選択の論拠となる，聴取音と擬音語表現の音響的関連性を概観する．続く第 4 章では第 3 章の議論に基づいて選択した特徴量を紹介し，第 5 章で本手法の性能を石原らによる従来法との比較実験により検討する．最後に第 6 章で結論と今後の展望を述べる．

2.　環境音の擬音語認識

擬音語認識における問題の一つとして，一つの音に対し想起される擬音語は聴取者によって変化する（一意に定まらない）という曖昧性がある．複数正解が存在しうるという点で，音声認識同様に通常の日本語音素で認識してしまうと曖昧性問題に対応できない．そこで石原ら [石原 05] は，特に曖昧性の強い音素を組合せた音声素のグループを設計し，認識時のクラスとして使用することで，システムの認識結果を一意に定めた．その結果，曖昧に認識されやすい音が入力された場合，その音に合った擬音語候補を複数出力するシステムが実現された．本稿で論じる環境音擬音語変換システムは本質的に従来法である上記システムの拡張に当たり，相違点は音素と音節構造認識の手法である．本手法の大まかな処理の流れ (図 1) を以下に示す．

1. 入力信号を時間領域でパワー包絡をチャンクに切り分ける (segmentation)
2. 各チャンクの音響特徴を抽出し，子音と母音をそれぞれ識別する (phoneme recognition)
3. チャンクの時間長や総数から音節構造を決定する (syllable structure determination)
4. 音素と音節構造を統合した擬音語を出力する (construction)

本稿では，ステップ 2,3 の音素認識における音響特徴抽出と，音節構造の決定方法を主軸にして議論する．

3.　聴取音と擬音語の音響的関連性

聴取音と日本語擬音表現の音響学的関係性は事例数は少ないが 16 年程前から報告され始めている [田中 95]．ここでは，中でも環境音の知覚とそこから想起される擬音語の，母音，子音，音節構造の決定に関する知見を概観する．

3.1　母音

比屋根ら [比屋根 98] はガンマトーンで模した衝突音の被験者聴取実験を行った．その中で刺激音の音響特性と聴取回答となる擬音語との関係が調査されたが，母音に関しては，信号の中心周波数が 1kHz 以下で /o/，1 〜 2kHz では /a/，2kHz 以上では /i/ と認識される結果となった．大石らは，純音 [大石 08] や周期的複合音 [大石 09] など比較的単純な音の擬音語表現に対応する周波数や音の特徴を検討しており，1/3 オクターヴ刻みで変化する純音を被験者に聴かせ，その音を擬音語で発声表現させ，擬音語表現が切り替わる周波数が求められた (188 Hz までは「ボー」，188−870 Hz までは「プー」，870 Hz 以上は「ピー」)．また，純音に倍音が加わった周期的複合音については，/i/ や /a/ の発音頻度が純音に比べて増え，/o/ と /u/ の使用頻度が減少することが確認された．このことから，信号のエネルギー分布の重心や帯域幅が高周波数側に移動することで擬音語母音の選択に変化が起こると考えられた．

より複雑な複合音と擬音語表現の関連性の特徴を調査するために，高田らは実際の環境音を刺激として 2 種類の聴取実験を行った [Takada 06]．1 つは被験者に呈示した刺激を擬音語で回答させるもので，回答を調音位置や母音など 24 個の音声学的特徴に符号化した．もう 1 つの実験では，刺激の聴取印象を音質の明瞭性や迫力など 13 の尺度を用いた SD 法で調査した．両実験から，擬音語系列の音声学的特徴と聴取印象に強い相関がみたれたものがあった．母音との関連に限定して述べれば，母音 /o/ は「暗い，鈍い，濁った」といった低い周波数帯域に主要なエネルギーを有する音と相関が高かった．一方で，「明るい，鋭い，澄んだ」といった高い周波数帯域に主要なエネルギーを有する刺激に対し，母音 /i/ が使われていた．/o/ と /i/ が使われた刺激音の平均スペクトル重心は，それぞれ 1593 Hz と 5308 Hz であった．日本人男性が発声する日本語 5 母音の平均スペクトル重心 (/o/: 1179 Hz, /u/: 1824 Hz, /a/: 1840 Hz, /e/: 2185 Hz, /i/: 2853 Hz) の順序と対応している．また，von Bismarck によると [Von Bismarck 74] ,「鋭さ」といった音の明瞭性に関する印象は，広帯域信号の，上限周波数，下限周波数，スペクトルの傾斜度合，が影響すると示唆されている．

以上の知見から，聴取音から想起される擬音語母音の決定には，信号の周波数軸上でのエネルギー分布，特に重心，帯域幅，偏り方などの特徴が強く関係していると考えられる．

3.2　子音

前節 3 件の研究からは子音の擬音語認知に関わる知見も得られている．比屋根らの実験 [比屋根 98] では，ガンマトーンに周波数ゆらぎを加えスペクトル形状を変化させることで，認識された擬音語が「かん」や「たん」という表現から，「ばん」「ぱん」「だん」などの濁音を含む表現に変化した．大石らによる周期的複合音の聴取実験では，倍音構造を持つ音に対して /b/ または /v/ などの濁音がつく子音が使われる傾向がみられた，

以上の知見から，聴取音と擬音語子音は，倍音構造などの詳細なスペクトル形状を表す，比較的高次な音色特徴と関連していると考えられる．

3.3　音節構造

比屋根ら [比屋根 98] は，促音，撥音，長音などの付属モーラの認識と単発音の残響時間の関連を調査した．ガンマトーンを用いた擬似単発音の認知実験により，信号の中心周波数が 4 kHz の条件では，残響時間が 0 〜 100 msec で「チッ（促音）」，100 〜 200 msec で「チン（撥音）」，300 msec 以上で「チーン (長音＋撥音)」として認識される傾向があると報告している．

4. 提案手法

4.1 チャンク分割

本稿では，入力信号を擬音語の 1 音節に対応するチャンクへ分割する方法に，石原らによる従来法で使われた手法を使う．各チャンクが擬音語音節に対応するという仮定は聞こえ度理論 [Ladefoged 00] に基づく．分割のルールは，音響波形のパワー包絡からピークと谷を検出し，隣り合うピークと谷の比が予め設定された閾値異常であれば，切り出してチャンクを生成する．

4.2 特徴抽出

4.2.1 母音特徴

3.1 節で述べた，スペクトル上でのエネルギー分布の重心，その帯域幅，そして偏り方を特徴として抽出するために，母音の特徴量にはそれぞれ，スペクトルの重心，分散，歪度を採用する．スペクトル重心 μ はパワースペクトル全体の平均値で，スペクトル分散 σ^2 は重心を中心としてどれくらいエネルギーが拡がっているか，即ちどの程度の帯域幅を持つかを抽象的に表現している．スペクトル歪度 γ は形状の非対称性を表す値であり，$\gamma < 0$ の場合，より高域側にエネルギーが偏り，反対に $\gamma > 0$ の時，低域側に偏る．上記 3 つ特徴量の算出式を以下に示す．

スペクトル重心 μ:

$$\mu = \frac{\sum_{n=0}^{N-1} f_n x_n}{\sum_{n=0}^{N-1} x_n} \tag{1}$$

スペクトル分散 σ^2:

$$\sigma^2 = \frac{\sum_{n=0}^{N-1} (f_n - \mu)^2 x_n}{\sum_{n=0}^{N-1} x_n} \tag{2}$$

スペクトル歪度 γ:

$$m = \frac{\sum_{n=0}^{N-1} (f_n - \mu)^3 x_n}{\sum_{n=0}^{N-1} x_n} \tag{3}$$

$$\gamma = \frac{m}{(\sigma^2)^{1.5}} \tag{4}$$

ここで N 周波数ビンのインデックス，x_n は各ビンのパワー，f_n は各ビンの中心周波数を表す．本稿では窓長を 4096 サンプルとし，これより長い入力信号はシフト長 2048 サンプルで窓分析しながら，全フレームで算出した値の平均値を特徴量とした．

4.2.2 子音特徴

子音特徴は，スペクトル包絡を効率的に表すことのできる MFCC として抽出する．後述の実験では，メルフィルタバンク数を 24 とし，17 次元 MFCC (0 次含む) + 17 次元 ΔMFCC を使用する．

4.3 音声素グループ決定

環境音に対する擬音語を音素レベルで擬音語を一意に定めるのは難しいため，擬音語表現として似た役割を持つ音素をグループ化した音声素グループ (表 1) [石原 05] レベルで識別を行う．例えば「ばん」「がん」「ぼん」「ごん」にも聞こえうる環境音は，子音の音声素グループは /b-g/ として認識し，母音音声素グループは /ao/ として認識する．本稿では，識別器として混合ガウスモデル (GMM) を使用する (混合数= 16)．

表 1　書き起こしに基づく音声素グループ

母音音素グループ
/ao/, /a/, /i/, /u/, /e/, /o/
子音音素グループ
/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/,

4.4 擬音語出力

4.3 節で決定した音声素グループに基づいて，一つ以上の擬音語を出力する．上述の例では，「ば」「が」「ぼ」「ご」に必要に応じて付属モーラ（促音・撥音・長音）を付加した擬音語を出力する．付属モーラの決定規則は， 3.3 で述べたように, チャンクの時間長が， 0 〜 100 msec で「ッ（促音)」， 100 〜 200 msec で「ン（撥音)」, 300 msec 以上で「ーン (長音＋撥音)」と付加するものとする．また各チャンクを 1 擬音語の 1 音節とみなしているため，各チャンクの音素認識を行った後，処理の最終段階でチャンクを結合して擬音語系列とした．

5. 評価実験

本稿での実験の目的は，心理音響などの分野から環境音知覚の知見を応用して選択した特徴量と音節構造決定手法の，擬音語認識性能を調査することが目的である．具体的には，

1. 子音と母音，それぞれの特性に合わせた特徴量を使用
2. パワー包絡の各チャンクの時間長と総チャンク数から音節構造を決定

上記二点が認識性能に与える影響を，被験者聴取実験を通じて従来法と比較しながら検討する．

5.1 実験条件

学習には RWCP の実環境音声・音響データベース [RWCP] から，上述した音声素グループ (表 1) で手動ラベル付けした 5,000 サンプルの非音声音ドライソース（環境音）を使用する．比較対象である従来法は，母音・子音・付属モーラ全てを同様に処理し，特徴量に 17 次元 MFCC + 17 次元 ΔMFCC を使い，3 状態, 16 混合の HMM で識別を行う．この時分析窓長は 75 msec，シフト長は 15 msec に設定した．

被験者実験では，被験者 2 人に評価用環境音を聴取し主観で擬音語ラベル付けをしてもらい，そのラベルを正解データと仮定して，同じデータセットを入力した擬音語認識システムの出力と照合し，再現率・適合率を算出した．評価データには, 効果音 CD [音源 1, 音源 2] から抽出した単発音 25 サンプルを用いた．

評価は従来法と同じく，システム出力結果の被験者回答に対する再現率と適合率に基づいて行う．複数正解が存在しうる擬音語認識では，再現率・適合率の計算方法が通常のものとは少し異なる．各評価尺度の定義式を以下に示す:

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

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

表 2　評価実験結果

手法	再現率	適合率
従来法	27/50　(54.0%)	16/23　(69.57%)
本手法	32/50　(64.0%)	18/27　(66.7%)

5.2　実験結果

実験の結果を表 2 に示す．再現率だけに関して言えば 10% 改善している，これは母音の認識精度の向上が貢献しており，母音に対しては，MFCC ではなく，より抽象的にエネルギーの分布を表現するような特徴量の使用が適切であることがわかった．顕著な例として「陶器を金属で軽く叩いた音 (聴感: ちーん，てぃーん)」があり，従来法では /t-ch u: N/ (つーん／ちゅーん) という母音だけ間違った擬音語が出力されるが，本手法では両被験者の聴感上と同じ /t-ch i: N/ が出力された．一方で適合率は約 3% 低下している．これは本手法によって表現できる音節構造の範囲が広まったが，必ずしも（被験者にとって）正確な擬音語を出力していないからだと言える．例えば「タイピング音 (聴感: かたかたっ)」があるが，現状の統合規則では /k-t a q/ + /k-t a q/ + /k-t a q/ + /k-t a q/ (例: かったっかったっ) , という不自然な出力が得られた．これは隣接するチャンク同士の間隔を考慮せずに音節結合をしてしまっているためであり，間隔が閾値より短ければ促音を省略するルールを設けることで解決できる．

6.　結論

本稿では，人間の環境音知覚及び擬音語認知研究の知見に基づいて選択した音響特徴を用いて，環境音の音響信号を擬音語として認識するシステムを設計した．具体的には:

1. 擬音語音素の認識段階で，子音と母音それぞれの特性に合わせて設計した別々の音響特徴量を使用
2. 切り出したパワー包絡の各チャンクの時間長と，チャンク数から音節構造を決定

の 2 点を行う新しい手法を使った擬音語認識システムである．システムの性能を評価するために，従来システムと擬音語認識結果を比較した．結果から，母音認識性能の改善が認められたが，音節構造の決定手法による誤認識が問題となった．今後の課題として，(1) 音節結合ルールの再検討，(2) 繰返し音に対する音声構造の決定手法，(3) 変調の周期などの動的な特徴量が音の印象に影響をあたえること知られているので [Fastl 07] 新しく特徴量として加えることを検討している．また，発展的な課題としては，音源の種類によって使用される擬音語語彙が限定されると仮定し，音源同定の結果を援用して認識に使用する音声素グループを切り替えるシステムなどが考えられる．

謝辞　本研究は科研費 (S)(A), GCOE 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

[Chan 10] Chan, C.-F. and Yu, E. W. M.: An Abnormal Sound Detection and Classification System for Surveillance Applications, *Signal Processing*, pp. 1851–1855 (2010)

[Chu 09] Chu, S., Narayanan, S., and Kuo, C.: Environmental Sound Recognition With Time-Frequency Audio Features, *IEEE Trans. on ASLP*, Vol. 17, No. 6, pp. 1142–1158 (2009)

[Cowling 03] Cowling, M.: Comparison of techniques for environmental sound recognition, *Pattern Recognition Letters*, Vol. 24, No. 15, pp. 2895–2907 (2003)

[Eronen 06] Eronen, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., Lorho, G., and Huopaniemi, J.: Audio-based context recognition, *IEEE Trans. on ASLP*, Vol. 14, No. 1, pp. 321–329 (2006)

[Fastl 07] Fastl, H., Zwicker, E.: Psychoacoustics: facts and models, Springer-Verlog New York (2007)

[Ladefoged 00] Ladefoged, P.: *A Course in Phonetics*, Thomson Learning (2000)

[Ntalampiras 10] Ntalampiras, S., Potamitis, I., and Fakotakis, N.: A Multidomain Approach for Automatic Home Environmental Sound Classification, in *INTERSPEECH-10*, September, pp. 2210–2213, ISCA (2010)

[Takada 06] Takada, M., Tanaka, K., and Iwamiya, S.-i.: Relationships between auditory impressions and onomatopoeic features for environmental sounds, *Acoustical Science and Technology*, Vol. 27, No. 2, pp. 67–79 (2006)

[Von Bismarck 74] Von Bismarck, G.: Sharpness as an attribute of the timbre of steady sounds, *Acustica*, Vol. 30, No. 3, pp. 159–172 (1974)

[Wake 01] Wake, S. and Asahi, T.: Sound Retrieval with Intuitive Verbal Descriptions, *IEICE Trans. on Info. and Sys.*, Vol. 84, No. 11, pp. 1568–1576 (2001)

[音源 1] 音源 1 キングレコード：効果音全集

[音源 2] 音源 2 キングレコード：新・効果音全集

[山内 03] 山内 勝也, 高田 正幸, 岩宮 眞一郎：サイン音の機能イメージと擬音語表現, 日本音響学会誌, Vol. 59, No. 4, pp. 192–202 (2003)

[RWCP] 実環境音声・音響データベース：RWCP Sound Scene Database in Real Acoustical Environments, http://tosa.mri.co.jp/sounddb/index.htm

[石原 05] 石原 一志, 駒谷 和範, 尾形 哲也, 奥乃 博：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, 人工知能学会論文誌, Vol. 20, pp. 229–236 (2005)

[大石 08] 大石 弥幸, 三品 善昭, 龍田 建次：純音を表す擬音語の周波数による変化ー年齢の影響ー, 日本音響学会 2008 年秋季研究発表会講演論文集, 3-7-2 (2008)

[大石 09] 大石 弥幸, 三品 善昭, 龍田 建次：周波数変調音を表す擬音語, 日本音響学会 2009 秋季研究発表会講演論文集, 1-5-5 (2009)

[田中 95] 田中 基八郎, 松原 謙一郎, 佐藤 太一：異音の表現における擬音語の検討： 衝突音等の単発音やうなり音の場合, 日本機械学會論文集. C 編, Vol. 61, No. 592, pp. 4730–4735 (1995)

[比屋根 98] 比屋根 一雄, 澤部 直太, 飯尾 淳：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. SP, 音声, Vol. 97, No. 586, pp. 65–72 (1998)

[筧 93] 筧 壽雄, 田守 育啓：オノマトピア: 擬音・擬態語の楽園, 勁草書房 (1993)