日本語の擬音語表現に着目した環境音自動認識

石原 一志† 坪田 康‡ 奥乃 博‡‡
†京都大学工学部情報学科 ‡京都大学学術情報メディアセンター ‡‡京都大学情報学研究科

1. はじめに

我々は様々な環境音を擬音語で表現し, コミュニケーションを行っている. このような擬音語認識が自動化できれば, 音データの忠実なアノテーションが可能となり, 音データやマルチメディアデータのより高度なアーカイブや検索が実現できるようになる. また, 字幕に『犬の鳴き声』ではなく『うぉーん』と出すことにより中途失聴障碍者らに自然な音情報を伝えたり, 音声認識の OOV (out-of-vocaburary) 問題に擬音語で対応するなど, より人間らしいヒューマンインターフェースの構築が可能となる.

我々は, 音声認識との対比から, 擬音語自動認識は 3 段階の処理過程を取ると考える. 図 1 に示したように, (1) セグメンテーションにより環境音を音節単位で切り分け, (2) 構造認識により「**モーラ構造**」を同定し, (3) 音素認識によりモーラ構造を擬音語に変換する. ここで, モーラ構造とは, S (撥音以外の音節), R (長音), Q (促音), N (撥音) から構成される系列である. 日本語における 1 モーラ語基の擬音語の音韻形態を表 1 に示す[1].

擬音語認識では, 環境音を擬音語で表現するときの標記法が聴者により異なるため, 「正しい」認識を定義することが難しい. 我々は, 「音節数レベルにおける認識は同一母国語の聴者ならばほぼ一致している」という『**音節数不変仮説**』を立てることにより本問題点を回避し, 聴者に依存しないレベルでの妥当な表現を自動認識することを目指す.

環境音データベース, とくに, 擬音語のラベルのついたデータベースは現在のところ整備されていないので, モーラ構造や音素レベルの認識のために信頼性の高い環境音認識用 HMM を構築することは難しい. 実際, 信頼性の低い環境音 HMM による認識結果や, 音声を用いて作成された音響モデルによる環境音の認識結果は, 音節数レベルにおいても人間の認識と大きく異なった. したがって, HMM のような統計的な手法ではなく, 環境音の種類によらない特徴量を用いて単純な規則からモーラ構造を決定する手法を使用する. 本稿では, 音節数不変仮説の検証と, この仮説に基づいた環境音の擬音語認識手法について報告する.

2. 音節数不変仮説の聴取実験による検証

音節数不変仮説を検証する目的で, 聴取実験として, 20 名の日本人被験者に対し動物の鳴き声や楽器音などの環境音[S1),S2),S3)]を聴かせ, 詳細な書き起こしを求めた. 被験者には各環境音が何の音であるかについての情報は与えていない. 22 サンプルの実験に対して, 約 7 割の音節数が一致し, さらに波形包絡のピーク数とも一致した. この聴取実験から本仮説は妥当であると判断を下した.

Enveronmental Sound Recognition using Japanese Onomatopoeia *by Kazushi Ishihara, Yasushi Tsubota and Hiroshi G. Okuno (Kyoto Univ.)*

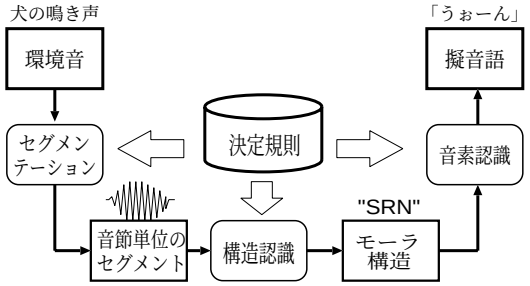

図 1 擬音語認識の処理過程

表 1 1 モーラ語基の擬音語表現

モーラ構造	擬音語例	モーラ構造	擬音語例
S	ふ	S R	さー, すー
S Q	きっ, さっ	S R Q	ざーっ, どーっ
S N	かん, ぴん	S R N	がーん, ごーん

表 2 セグメンテーション比較実験 (再現率/適合率 (%))

ピーク単位	ASR (Julian)	環境音 HMM
83.7 / 99.1	100.0 / 26.2	89.1 / 38.9

3. 環境音の擬音語認識

3.1 音節単位へのセグメンテーション

前章の聴取実験から得られた「環境音に対する書き起こしの音節数は波形包絡のピークの数にほぼ一致する」という知見を用いて音節単位を切り出す. 具体的には, 波形包絡における 2 つのピーク間の極小値と小さい方のピークのパワー比を求め, その値が閾値以下である全ての極小点で切り出して, ピーク単位でのセグメントをとる.

ピーク単位でのセグメンテーション結果を, 音声認識システム (Julian) での切り出し, 環境音 HMM による切り出し結果と比較する. 環境音 HMM は, 動物の発声などの環境音 3795 サンプル[S4),S5)] を 6 回学習させて構築した混合数 16 のモノフォンモデルである. 特徴量として 26 次元 MFCC を用いた. また, 学習サンプルには 3.3 節で後述する音素グループレベルでラベルを与えている. 無作為に選択した環境音のセグメント位置 157 点において 3 システムのセグメンテーションを比較した結果を表 2 に示す. これによりピーク単位でのセグメンテーション手法が有効であることが分かる.

本比較実験は, 各セグメントに 1 音節からなるモーラ構造を仮定した. 実際には, 日本語の文法に則した表現, あるいは音響的類似性よりもその音が持つ特徴を優先した表現のために, 聴者が無意識に存在しない音素を加えることがあり, 一部のセグメントは聴者によっては複数の音節に認識される[2]. しかし, 本実験は聴者の表現方法に依存しないレベルでの認識を目的とするため, これらの事象の影響を受けないラベルを用いて行った.

表 3 促音・撥音の有無決定に用いた特徴量

ピーク点からピークの 5 %以下のパワーになる点 (Min) までの
1. 音長, 2. 平均傾き, 3. 最大傾き, 4. 最大傾き後の音長
5. 各点と Min を結んだ直線のうち最大の傾き

表 4 モーラ構造認識の比較実験結果 (再現率/適合率 (%))

	C5.0	ASR (Julian)	環境音 HMM
促音	82.9 / 66.5	7.1 / 33.3	21.4 / 75.0
撥音	22.7 / 35.7	9.1 / 33.3	9.1 / 100.0
長音	84.6 / 100.0	40.0 / 100.0	66.7 / 71.4

3.2 モーラ構造の認識

前節で切り出されたセグメントからモーラ構造を認識する. 各セグメントの構造は, 日本語で用いられる 1 モーラ語基の擬音語パターン (表 1) から “ S [R] [N] [Q] ” という文法で定義できる☆[1)]. すなわち, 長音, 撥音, 促音の有無を決定すれば, モーラ構造は一意に定まる. 聴取実験で用いた環境音の波形とスペクトログラムから, ピーク後の減衰速度が大きい音は促音に, 減衰速度が小さく長時間減衰が続く音は撥音に認識されやすいという傾向が見られた. 一方, 長音 (R) は音長に依存しており, 聴取実験において半数以上の被験者の書き起こしに長音が現れたサンプルの割合は, 400ms 以上のサンプル集合においては 8 割以上であったのに対し, 400ms 以下のサンプル集合には全く存在しなかった.

これらの知見は, 単発音の先行研究でも報告されている. 単発音は 1 モーラ語基の擬音語で表されるという点で環境音から切り出したセグメントに近い. 田中らは, 残響がしばらく続く単発音は撥音を付けて表すことが多いと報告している[3)]. また比屋根らはガンマトーンを用いた疑似単発音に対し, 4kHz では 0〜100ms がチッ, 100〜200ms がチン, 300ms 以上がチーンと表現されるという実験結果を報告している[4)]. これらの報告は「減衰速度や残響, 音長が促音や撥音, 長音の有無を決定する」という上記の聴取実験からの予測と一致している.

このことをふまえて特徴量を設計し, その中から選定した特徴量 (表 3) を用いて評価用サンプルの促音・撥音の有無を決定した. 評価には聴取実験で用いたサンプルを, 4 分の 1 以上の被験者が促音 (撥音/長音) を認識したサンプルとそれ以外のサンプルに分けて用いた. また, 認識手法として, 分岐 3 回までに限定した C5.0, Kernel 線形および 2 次の SVM, k-NN 法を検討し, 最も有効であった C5.0 を本実験では用いた. 統計的手法である音声認識システムによる認識, 環境音 HMM による認識との比較結果を表 4 に示す. これにより当手法はモーラ構造認識において, 音声認識システムや環境音 HMM による認識に比べて有効であることが言えた. しかし, 全体的に適合率・再現率共に低く, 特に促音・撥音においては現段階では認識手法としての有効性を言うことはできない. 人間の認識では, 音長や減衰速度の他に周波数的な特徴や音源の種類による影響なども関係していると考えられるため, 今後より適切な特徴量を検討していく. また, セグメントの統合後に行う語の補正 (整語) の段階で, 促音や撥音の有無を決定する手法についても検討する.

☆ 角括弧は省略可を表す

表 5 音素グループ

1. m n my ny (ナ, マ行 (鼻音)) 2. s sh z (サ, ザ行 (摩擦音))
3. h f hy (ハ行 (摩擦音)) 4. w j (ア, ヤ, ワ行)
5. b d g by dy gy (ガ, ダ, バ行 (有声破裂音)) 6. r (ラ行 (流音))
7. p t k py ts ch py (カ, タ, パ行 (無声破裂音))

3.3 音素グループの自動構成

前節で得たモーラ構造を擬音語に変換するための音素認識手法を検討する. 音素レベルにおける認識結果は聴者に依存し一意に定まらないが, 認識されやすい音素の傾向はあると考えられる. そこで, 聴者に依存しないトランスクリプションレベルとして類似した音素からなる「**音素グループ**」を設計する. これは, 聴取実験結果を用いて, 同じ環境音に対して頻繁に認識される音素同士を類似した音素としてクラスタリングすることにより作成する. その結果, 表 5 に示す音素グループ分類が得られた. 3.1 節および 3.2 節の実験で用いた環境音 HMM は表 5 の音素グループレベルでラベルをつけた学習サンプルを用いて構築している.

この分類結果は, 鼻音・摩擦音・破裂音などといった調音方法による分類とほぼ一致している. これは, 同じ調音方法を持つ音素は類似した環境音の特徴を表現するのに用いられやすいということであり, このことは言語学における音象徴の現象に一致する. 音象徴とは音素そのものが語が持つ意味とは別に象徴的な意味を示唆することである[1)].

今後, この音素グループを用いて環境音の認識実験を行い, その分析結果からより適切な音素グループを設計する. しかし, 音素分類が音源の種類に依存して複数存在することも考えられるため, 環境音情報に応じて様々な音素分類を調査する必要がある.

4. おわりに

音節数不変仮説の妥当性を検証し, これに基づいてモーラ構造を利用した環境音認識手法を検討した. 今後の課題として, 音素認識手法の精細化, 認識された各セグメントの統合, および統合した語の整語について検討を行う. さらに音源の種類による影響や言語・文化などからうけるバイアスの調査, 聴覚心理学の研究の利用などを行い, 最終的にこれらを統合して自然な擬音語出力を行うシステムを構築する. なお, 本研究は電気通信普及財団の研究助成を受けた.

参考文献

1) 田守育啓: 『オノマトペ – 形態と意味 –』, くろしお出版, 1999.
2) 田中基八郎: 機械の異常音の擬音語表現, 日本音響学会誌, Vol.53, No.6, pp.477–482, 1997.
3) 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.
4) 比屋根一雄 他: 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, 信学技報, SP97-125, 1998.

[S1] 蒲谷鶴彦：日本野鳥大鑑増補版 鳴き声 420, 小学館.
[S2] 新効果音大全集,KING RECORD.
[S3] JUST BIRDS&ANIMALS, テックシステム.
[S4] ANIMAL TRAX, クリプトン・フューチャー・メディア.
[S5] WORLD WIDE OF ANIMALS, クリプトン・フューチャー・メディア.