3L-7

擬音語表現を利用した 環境音のためのXMLタグの設計と自動付与

田口 明裕 † 北原 鉄朗 ‡ 石原 一志 ‡ 駒谷 和範 ‡ 尾形 哲也 ‡ 奥乃 博 ‡

† 京都大学 工学部情報学科 ‡ 京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

計算機と人間が様々な音に対して注意を共有し，インタラクションに取り入れていくには，音を計算機と人間の双方が理解できる形で表現できなければならない．音声には言語表現，音楽には楽譜表現があるが，それ以外の音（環境音と呼ぶ）には確立した表現法は今のところ存在しない．

一方，芦谷らによる鳥の鳴き声認識 [1] など，計算機上で環境音を扱った研究は少数ながら存在する．しかし，それらの手法は環境音の音源同定を目的としており，音源が既知・明確である音を分類するに留まっていた．

我々は,「何の音か」ではなく「どう聞こえるのか」の観点から環境音を記述する重要性を指摘し，環境音を擬音語として記述するシステムを実現した [2][3]．これにより，音源の分からないような音に対しても擬音語を通じて情報の共有が可能になった．しかし，基本的に単セグメントの音のみを扱っており，複数のセグメントが連なるような音は扱っていなかった．複数セグメントの音は，音高の変化やリズムといった情報を用いることができるので，よりリッチな表現が可能である．本稿では，擬音語・音高・リズムの3つを用いて複数セグメントの音を表現するXMLタグとその自動付与法を提案する．

2. 環境音のためのXMLタグの設計

環境音のためのXMLタグの設計における我々の方針は,「人間による音の口まねを過不足なく表現できる」ことである．人は，どのように聞こえる音かを表現するときに通常口まねを行う（図1）．もちろん，人間が音を完全に模倣するのは不可能であり，お互いに理解できる範囲で単純化がなされるのが一般的である．我々は，さまざまな音に対する口まねを分析した結果，次のような傾向を得た．

1. 音が切れて聞こえるところは切って発音する．本稿では，切って発音された各々をセグメントと呼ぶ．これは，実際の音の切れ目以外にも音高や音色が急激に変化するところでも発生する．
2. 音高は，絶対的な高さよりも相対変化が重視される．相対変化はセグメント間とセグメント内の両方が表現される．
3. セグメントの時間的構造（リズム）は比較的正確に表現される．
4. 音色は主に擬音語によって表現される．

以上から，音をセグメント列として表現し，各セグメントおよびセグメント間の関係を音高，リズム，擬音語によって表現する．以下，この3要素の各々について設計方針を論ずる．

Design of XML Tagset for Environmental Sounds based on Sound-imitation Words and its Automatic Annotation

by Akihiro Taguchi, Tetsuro Kitahara, Kazushi Ishihara, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)

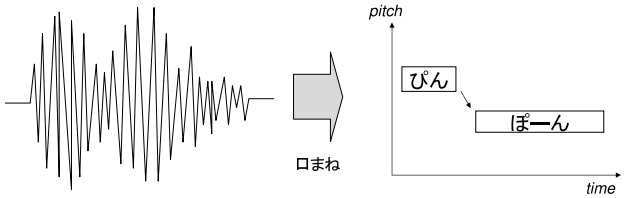

図 1: 環境音の口まね

表 1: 調音法に基づく音素グループ

音素グループ	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/	有声破裂音
/u-exp/	/ch/, /k/, /p/, /t/, /ts/	無声破裂音

• **音高**　前述の通り，セグメント間およびセグメント内の音高の相対的な時間変化を表すことが重要である．また，表現の粒度は，特徴を十分に表現できる範囲内で個人差が出ない程度に荒いことが望ましい．そこで，セグメント内の音高変化は「一定・上昇・下降・山型・谷型」の5種類から，隣接セグメント間の音高変化は「一定・上昇・下降」の3種類から選択する．

• **リズム**　リズムを捉えるために重要となるのは，セグメントの発音開始時刻，発音時間，そして発音終了から次の発音までの間隔が挙げられる．そのため，各セグメントのオンセット及びオフセットの記述を行う．

• **擬音語**　擬音語は多種多様であり，聴取者などによって表現が変わってしまうというあいまい性が生じる．このあいまい性をできるだけ抑えるため，まず擬音語構造に制限を加える．1セグメント1音節とし，1音節を「子音 母音 促音—撥音」とする．さらに，子音を音素レベルで表現するのではなく，調音法に基づいた音素グループレベル (表1) で表現する．

以上の方針に基づいて設計したXMLタグセットを表2に示す．

3. 自動付与手法

前節で設計したタグ表現を音響信号から自動的に得る手法について述べる．処理の流れを図2に示す．

3.1 セグメンテーション

パワー包絡によって音の切れ目を検出し，分割を行う．しかし，パワーの大きな変動が見られない場合でも，音色や音高の急激な変化によっても擬音語の変化が起こりうる．このような擬音語変化にも対応するため，一度パワー包絡を利用して分割した後，スペクトル変形の観点からセグメントを再分割する．まず，音響信号にフィルタバンクを適用して時間・周波数平面におけるパワーの分布を得る．次に，時刻ごとにパワーが閾値以上のバン

表 2: 設計したタグセット

タグ	子要素	属性	意味	取りうる値
<Sound>		locate	Wav ファイルの参照位置	文字列
		length	音長 (sec)	正実数
		segment_num	セグメント数	正整数
	<Time> <SIW> <Pitch>	id	セグメント番号	正整数
<Time>		onset	開始時刻 (sec)	正実数
		offset	終了時刻 (sec)	正実数
<SIW>		C	擬音語（子音）	表 1 参照
		V	擬音語（母音）	'a', 'i', 'u', 'e', 'o'
		QN	擬音語（促音 or 撥音）	'Q', 'N'
<Pitch>		Reliability	音高情報信頼度	実数
		Outside	前セグメントとの音高推移	'flat', 'up', 'down'
		Inside	セグメント内の音高推移	'flat', 'up', 'down', 'peak', 'trough'

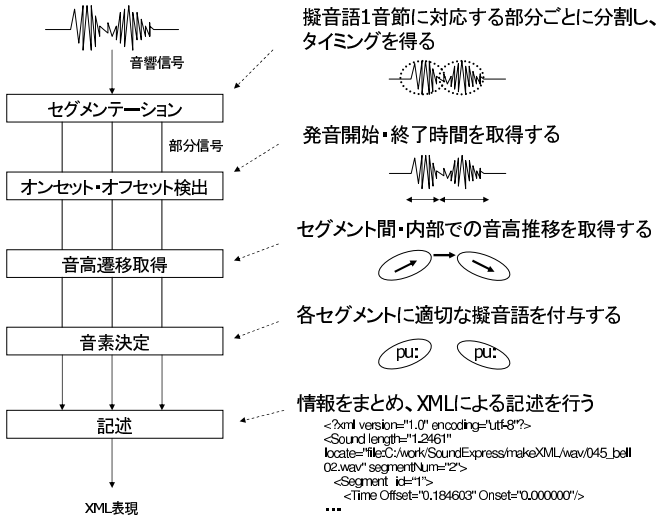

図 2: 処理の流れ

クを抽出し，時間方向へ同一あるいは隣接バンクを繋いでゆく．そして，この切れ目となったところでセグメントを分割する．

3.2 オンセット・オフセットの検出

各セグメントに対して，Klapuri の手法 [4] でオンセットを検出する．オフセットは次セグメントとの境界時刻とする．

3.3 音高推移パターンの取得

環境音を扱う際には，調波構造を持たない音が多く現れるため，基本周波数を推定する手法では音高を正しく比較することはできない．そこで我々は，次のようにして音高遷移パターンを取得する．まず，各セグメントを時間方向に序盤・中盤・終盤の 3 分割し，それぞれのパワーが最大となるバンクの周波数を計算する．そして，この 3 つの周波数から「上昇」「下降」「山なり」「谷形」「一定」のセグメント内の推移パターンを判定する．セグメント間の遷移パターンは，前セグメントの終盤の周波数と後セグメントの序盤の周波数を比較して判定する．ただし，特定のバンクのパワーが十分に大きくならないような音に対しては，信頼できる結果が得られないので，最大パワーのバンクのパワーの優勢さに基づいて信頼度を計算する．

```
<?xml version="1.0" encoding="utf-8"?>
<Sound length="0.9434"
  locate="file:C:/makeXML/wav/042_beep14.wav"
  segmentNum="2">
  <Segment id="1">
    <Time Offset="0.114762" Onset="0.000000"/>
    <SIW C="u-exp" QN="Q" V="ao"/>
    <Pitch Inside="flat" Outside="none" Reliability="0.000363"/>
  </Segment>
  <Segment id="2">
    <Time Offset="0.957846" Onset="0.154626"/>
    <SIW C="v-exp" QN="Q" V="ao"/>
    <Pitch Inside="flat" Outside="up" Reliability="0.002156"/>
  </Segment>
</Sound>
```

図 3: XML タグ例「クイズの不正解時に流れるブザー音」

```
<?xml version="1.0" encoding="utf-8"?>
<Sound length="0.90866" locate="file:C:/makeXML/wav/058_door_13.wav"
  segmentNum="4">
  <Segment id="1">
    <Time Offset="0.194580" Onset="0.000000"/>
    <SIW C="u-exp" QN="Q" V="ao"/>
    <Pitch Inside="trough" Outside="none" Reliability="0.062374"/>
  </Segment id="2">
  <Segment>
    <Time Offset="0.444014" Onset="0.229456"/>
    <SIW C="v-exp" QN="Q" V="ao"/>
    <Pitch Inside="trough" Outside="flat" Reliability="0.043027"/>
  </Segment id="3">
  <Segment>
    <Time Offset="0.678481" Onset="0.458934"/>
    <SIW C="v-exp" QN="Q" V="ao"/>
    <Pitch Inside="trough" Outside="up" Reliability="0.030113"/>
  </Segment id="4">
  <Segment>
    <Time Offset="0.883016" Onset="0.678435"/>
    <SIW C="u-exp" QN="Q" V="ao"/>
    <Pitch Inside="down" Outside="up" Reliability="0.000000"/>
  </Segment>
</Sound>
```

図 4: XML タグ例「ドアを開け閉めしたときの音」

3.4 音素決定

音声認識の分野で一般的に利用されている隠れマルコフモデルを用いて音素を決定する．特徴量には MFCC（16 次元）+パワー+ΔMFCC（16 次元）+Δ パワーを用いる．

4. 環境音からの XML タグ付与例

本手法を用いて，「クイズで不正解時に流れるブザー音」と「ドアを開け閉めしたときの音」から XML タグを自動付与した例を図 3，図 4 に示す．いずれの音も正しくセグメンテーションされ，音高や擬音語なども妥当なものを選ぶことができた．

5. おわりに

環境音のための XML タグを設計し，自動付与法を提案した．今後，環境音の検索などに応用することで，本タグの有効性を確認する．

謝辞 本研究の一部は，科研費，21 世紀 COE の支援をうけた．

参考文献

[1] 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”, 信学技報, SP92-13，1992.
[2] 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”, 人工知能学会論文誌 20, 3, pp.229-235，2005.
[3] 石原一志 他: “環境音の擬音語変換における音素決定曖昧性の解消”, 第 67 回情処全大，pp.285-286，2005.
[4] Anssi Klapuri, “Sound onset detection by applying psychoacoustic knowledge”, *ICASSP-1999*, pp.3089-3092.