同時複数音源に対する擬音語による音源選択システム

山村 祐介 † **高橋 徹** ‡ **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

実世界において人間と計算機が音環境を共有するには, 音に対する一意な参照が必要であるが, 複数の音が混在する環境ではそのような参照は困難である. そこで, 音に擬音語, 例えば打撃音に“コンコン”という文字列を与え, 音の相互参照を可能とする.

本稿では, ユーザが環境音の擬音語表現を入力すると, その音源の情報を提示する, 音源選択システムについて述べる. 擬音語による音源選択は, Schniderman の提唱した *“Visual Information-Seeking Mantra”* [1] における Zoom and Filter 操作の, 新しいインタフェースと見做すことができる. また従来の音可視化システム [2] は音声のみを扱っており, 環境音に対応していない. 環境音には音源名の分からない未知音源が多く, 擬音語を用いた未知音源の可視化により, 取り扱える音源の範囲が大きく広がる.

擬音語クエリによる音源選択の課題は, 同一音源に付与される擬音語がユーザ毎に異なり, システムと共通の擬音語を得られるとは限らないことである. システムとユーザの擬音語間の差異を埋め, 音源を一意に参照可能とする仕組みが必要となる.

2. 擬音語クエリによる音源選択

本システムは混合音とユーザクエリ（擬音語テキスト）を入力とし, ユーザの指定音源と類似性の高い音源の音源情報をランキングで提示する. ここで, 音源情報は音源と音源の到来方向に限定する. 具体的には図 1 で示すような 3 段階の処理を通して実現される.

フェーズ 1： 混合音の音源定位 · 分離
フェーズ 2： 各分離音の擬音語への変換
フェーズ 3： 分離音の音源情報のランキング表示

本システムでは次の 3 つを仮定する. (1) 同時に混在する音源数は最大 3, (2) 入力は短時間で減衰する音 (単発音), (3) 擬音語の変換結果は, /子音/+/母音/+/語末/. (3) について, 語末は促音, 撥音のことであり, 擬音語の文法に制約を設けたのは, 文語中に現れる擬音語は一般にこの仮定が成り立つという考えに基づいている [3].

フェーズ 1, 2 は従来手法 [4, 5] で解決するため, 本研究ではフェーズ 3 に焦点を当てる. フェーズ 3 までの処理を以下に述べる.

2.1 フェーズ 1: 混合音の音源定位・分離

音源定位 · 分離には, ロボット聴覚ソフトウェアの HARK [4] を利用する. HARK のモジュールとして実装されている MUSIC 法, 幾何制約付き高次無相関化音源分離法 (GHDSS) により, 前者で音源定位後, 得られた幾何情報を元に後者で音源分離を行う.

2.2 フェーズ 2: 各分離音の擬音語への変換

分離音の擬音語変換には, 石原らの手法 [5] を利用する. この手法では, 環境音信号を HMM で認識し擬音語に変換する. 擬音語表現が聴取者に依存することを考慮し, 環境音用の音素の集合を設計し, それに基づいてコーパスに音素ラベルを割り当てている.

音源分離における分離歪みにより, 擬音語変換の変換ミスが誘発される. 例えば“ピーッ”という笛の音が他の分離音に混入し, 変換後の擬音語に音素“p”が現れることがある. これに対して, 音響モデルの学習データにフェーズ 1 の方法で分離した音を加え, 分離歪みによる影響を緩和する.

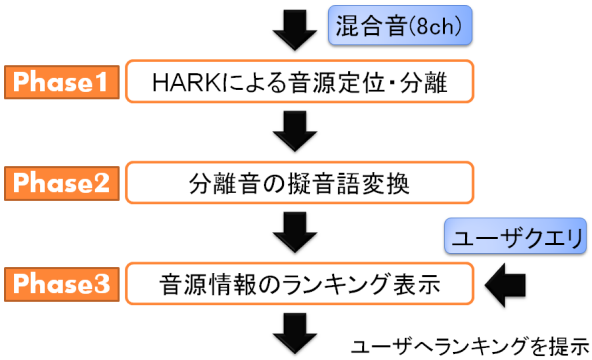

図 1: 擬音語クエリによる音源選択システムのイメージ図

3. 分離音の音源情報のランキング表示

本フェーズでは, 擬音語間の類似度を設計し, それを用いてユーザクエリに近い順に分離音の音源情報を提示する. 類似度設計の課題は次の 2 つである. (1) 厳密な一致判定では, 擬音語表現の揺らぎに対処できない, (2) 音源毎に明確な差が出ないと, 複数の音源が同様の類似度を持ち, 1 つの音源を選択できない.

問題設定は以下の通りである.

入力： 分離音の擬音語音素列
ユーザクエリ（擬音語テキスト）
出力： 分離音の音源情報のランキング表示

ランキングはユーザクエリと類似度の高い音源から順に 3-best でその情報を提示する. システムの出力をランキング表示にすることで, 最終的な判断をユーザに委ねる方法を取っている.

3.1 擬音語類似度の設計

MED(最小編集距離) を元に類似度を設計する. MED には挿入, 削除, 置換の 3 種類のコストがあり, 各変換コストを I, D, S と記述する. ここで, 基本となる MED (Basic MED) のコストを, $I = D = 1, S = 2$ とする. 本手法では, $I = D = 1$ とし, S は音素毎に異なるコストを付与する.

S は, フェーズ 2 で作成した音響モデルから, 2 つの音素の確率分布間の Kullback-Leibler 情報量 (KLD) により決定する. 音素の確率分布は 16 混合 34 次元の GMM で表現される. 音素を子音, 母音, 語末という 3 クラスに分け, 2 つの音素が別のクラスであれば, 置換コストを ∞ とする. 2 つの音素 p, q が同じクラスであれば, それぞれの音素の確率分布を P, Q とし, 置換コスト $S(p, q)$ を以下の

Sound-sources Selecting System by Using Onomatopoeia Querries from Multiple Sound-sources: *Yusuke Yamamura, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

式で定義する．

$$S(p,q) = KL_{SYM}(P,Q) * \frac{I+D}{KL_{MAX}}$$

ただし，$KL_{SYM}(P,Q)$ は確率分布 P,Q の間の KLD の平均である．

$$KL_{SYM}(p,q) = \frac{KL(P||Q)+KL(Q||P)}{2}$$

双方向から KLD の平均を取ることで，距離の対称性を満たしている．KL_{MAX} は KL_{SYM} の同クラス内における最大値であり，$0 \leq S(p,q) \leq I+D$ を満たすよう $S(p,q)$ を正規化する．

4. 実験

設計した類似度による音源選択の性能を評価する．

4.1 実験設定

RWCP 実環境音・音響データベース [6] の非音声ドライソースから単発音を 4,287 ファイル使用する．音源はすべて 16bit/16kHz でサンプリングされている．全ファイルには予め 1 人の手で正解ラベルが付けられており，これを 9:1 に分割し，各々学習用データ，評価用データとする．評価用データには，被験者 5 名により新しく正解ラベルを付与した．混合音は HRP2 (図 3) で実測した (0 °, 60 °, 300 °) のインパルス応答を畳み込み，足すことで生成する．

システムの出力でランキング 1 位の音源が，ユーザの指定音源と一致した割合を正解率とし，これを評価基準とする．ランキング 1 位が複数ある場合，ランダムに 1 位を選択する．

4.2 実験手順

実験は，混合音の入力から正解の判定まで (図 2) を試行 1 回とし，これを 500 回繰り返す．

入力となる混合音は，評価用データからランダムに選択した 3 種類の音源を用いて混合して作成する．混合音作成に使用した 3 種類の音源からランダムに 1 音源選択し，その音源に付与されている正解ラベルをユーザクエリとする．

4.3 実験条件

実験条件を，(A)Basic MED, (B) 本手法の MED, (C) 本手法の MED + 非混合，と変化させた場合の正解率を比較する．(A), (B) は 4.2 の実験手順に沿い，類似度判定に各々の MED を用いる．(C) は本手法の MED の使用に加え，混合と分離をせず，ランダムに選択された 3 音源からの選択を行う．

4.4 実験結果

条件毎の正解率を表 1 にまとめる．置換コストを音素毎に設計することで，Basic MED よりも正解率が向上した．しかし混合数 3 での音源選択のチャンスレートが 33.3%であることを考えると，正解率は十分高くない．この原因として次の 2 つが考えられる．

最大の要因として，本手法の類似度判定の学習データへの依存性が挙げられる．被験者 5 名が評価用データに与えたラベルは多様であり，本実験ではユーザの擬音語の多様性に対応できなかった．多くのユーザがラベル付けをした大量の学習データを用意し，それから作成した音響モデルで音素間の距離を求める必要がある．

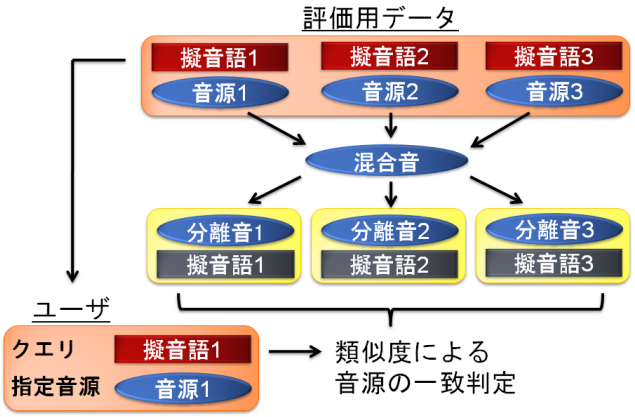

図 2: 実験手順 (試行 1 回)

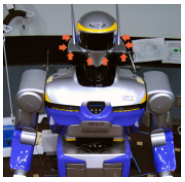

図 3: HRP2

実験条件	正解率
(A)	35.2%
(B)	48.2%
(C)	53.2%

表 1: 条件別の正解率

他に，漏れノイズによる擬音語変換誤りが多数あり，2.2 で述べた分離歪みへの対処以外にも対策が必要である．実験では (B) と (C) の正解率に余り差はないが，上で述べたデータへの依存性の問題を解決すれば，この問題は正解率に大きな影響を与えると予測される．

また今回は $I = D = 1$ としたが，音素毎の挿入と削除の設計も，正解率を向上させるアプローチになり得る．何故なら，ある環境音の擬音語に対して挿入或いは削除される音素は一様でない．例えば，“シャッ” というハサミで紙を裂く音は，人によっては “カシャッ” と表現されるが，“パシャッ” と表現されることはない．この場合，音素 “k” は音素 “p” よりも挿入されやすく，挿入コストを低く設計することができる．

5. おわりに

同時に存在する複数の音源からユーザが擬音語で指定した音源の情報を返すシステムの設計と，正解率による評価実験を行った．KLD により設計した距離尺度が従来の MED よりも正解率が高い一方で，全体的に正解率が低いことを確認した．今後は分離音の変換ミスや距離尺度の学習データ依存性への対処を行っていく．

謝辞 本研究の一部は科研費 (S), GCOE の援助を受けた．

参考文献

[1] B.Shneiderman: “Designing the User Interface (3rd Ed)”, Addison-Wesley, 1998.

[2] Y.Kubota, et al: “3D Auditory Scene Visualizer with Face Traking : Design and Implementation For Auditory Awareness Compensation”, doi:10.1109/ISUC.2008.59.

[3] 田守育啓: “オノマトペ-形態と意味-”, くろしお出版 (1999).

[4] K.Nakadai, et al: “Design and Implementation of Robot Audition System HARK Open Source Software for Listening to Three Simultaneous Speakers”, Advanced Robotics, Vol.5, No.6, pp.739–761, 2010.

[5] 石原一志, 他: “擬音語自動認識に基づいた環境音検索システム”, 情報処理学会 第 67 回全国大会, 4R-1, 2005.

[6] 実環境音・音響データベース: “Rwcp sound scene database in real acoustical environments”, http://tosa.mri.co.jp/sounddb/index.htm.