環境音から擬音語への自動変換における特徴量抽出法の検討

山川 暢英† 北原 鉄朗‡ 高橋 徹† 駒谷 和範† 尾形 哲也† 奥乃 博†

†京都大学大学院 情報学研究科 知能情報学専攻 ‡関西学院大学 理工学研究科／ JST CREST CrestMuse プロジェクト

1. はじめに

我々が日常の生活で知覚する音には，人の音声と音楽の他に，第3のカテゴリとして環境音がある．これをコンピュータに理解させることで，音情報による環境理解 (Computational Auditory Scene Analysis) や危険察知などへの応用が考えられる．しかしここで重要な課題は，「環境音をどのようなシンボルで記述するか」という表現法である．

その解決策として我々は，音源名ではなく「音がどのように鳴っているか」という情報のシンボルである擬音語として環境音を認識するシステムの研究開発に取り組んでいる．これまでに音源に依存しない環境音と音素グループへの自動変換が可能となっている [6]．しかし，環境音には繰り返し音，突発音，調波構造の有無など音響的特徴が多岐にわたっているので，音声認識で用いられるものとは異なった音響特徴量の設計や，音源同定の必要性が指摘されていた．

環境音認識のための特徴量の設計においては，突発音に代表される分析窓内での非定常性，あるいは，隣接窓間の時間変化を考慮することが不可欠である．前者については，音声認識でよく使用される MFCC (Mel-Frequency Cepstrum Coefficient) ではうまくとらえきれない．Chu ら [1] は非定常信号から音源同定を行う方法として，Matching-Pursuit (MP) [3] に Gabor 基底を用い，時間周波数領域で分析窓内の特徴を抽出することを提案した．しかし Chu らは，後述するような隣接窓間の時間変化は検討をしていない．また Gabor 基底で窓内の非定常成分を検出できていると主張していたものの，音源同定性能の点で他の基底関数と比較をしておらず，窓内定常性を仮定しないことによる効果を明確化していない．

本稿では，環境音認識により適した特徴抽出・同定手法を考察するために，分析窓内と隣接窓間の定常性を仮定する・しない場合をモデル化し，各仮定の妥当性を窓長を変えながら検討する．

2. 特徴抽出と時間変化のモデル化

分析窓をシフトさせながら音響信号の特徴抽出をする場合，非定常性 (即ち信号の時間変化) は“窓内”と“隣接窓間”に存在するので，両者は分けて扱う必要がある．

2.1 隣接窓間の時間変化

窓間の時間変化をモデル化する手法として，隠れマルコフモデル (HMM) が知られている．これにより信号の過渡特性を確率的にモデル化可能である．また HMM の状態数を 1 とすることで，ガウス混合モデル (GMM) を窓間の定常性を仮定した条件に適用できる．

2.2 分析窓内の時間変化

窓内の時間変化は，マザーウェーヴレットである Gabor 基底を窓内で伸縮・シフトし，局所的な変化を捉えることで検出できる．定常性を仮定する場合は，分析窓分の長さを持った正弦波 (Fourier 基底) を用いれば良い．またこの時，窓内でエネルギーの高い成分のみ検出し，意味のある変化を特徴として抽出するのが望ましい．

本稿では「解析に使う基底を自由に設定できる」，「エネルギーの高い成分を効率的に抽出できる」という利点から，Fourier / Gabor 基底両方を特徴抽出に使用できる枠組みとして MP をベースとする．

2.3 Matching-Pursuit による特徴抽出

MP は，所与の信号 s を，任意の m 個の基底信号 $\phi_{\gamma_1} \ldots \phi_{\gamma_m}$ の線形和として近似するアルゴリズムである：

$$s = \sum_{i=1}^{m} \alpha_{\gamma_i} \phi_{\gamma_i} + R^{(m)} \quad (1)$$

ここで $R^{(m)}$ は残差信号を表し，m 個の基底信号は $m'(\geq m)$ 個の基底信号が格納された基底辞書 $D = \{\phi_{\gamma_1} \ldots \phi_{\gamma_{m'}}\}$ から，次のようにして選択される：

1. D に含まれる各基底に対して s との相関を計算し，その値が最も高い基底を ϕ_1 としてその相関係数 α_{γ_1} と共に s から抽出する．
2. 残差信号 $R^{(1)} = s - \alpha_{\gamma_1} \phi_{\gamma_1}$ に対して 1. と同様の処理を行い，$\alpha_{\gamma_2} \phi_{\gamma_2}$ を得る．
3. 以上の処理を基底が任意の m 個抽出されるまで繰り返す．

本稿では Chu らに倣い，抽出基底の周波数と幅を特徴量とする．1 つの基底から 2 種類の特徴量が得られるので，特徴空間全体の次元は，抽出基底数×2 となる．

3. 同定性能の比較実験

3.1 実験条件

比較実験用の音源には，RWCP 実環境音・音響データベース [4] の非音声音ドライソースから，金属衝突音系 (貯金箱，コーヒー缶，コーラ缶，フライパン)，噴射音系 (ポンプ，スプレー)，楽器系 (ベル，カスタネット) の合計 8 音を用いた．これらは減衰の速い音 (金属衝突音) と遅い音 (その他) の 2 種類に分類できる．音源は全て 16bit/16kHz でサンプリングされている．それぞれ同じ音源で発音方法を微妙に変えながら録音したものが 100 個用意されている．

窓内の定常性を仮定することによる影響を検証するために，Fourier 辞書と Gabor 辞書を用意しそれぞれの同定性能を比較する．窓内の定常性を仮定することの妥当性は，窓長によって変わると考えられるので，20msec と 100msec の 2 種類の窓長に対して実験を行う (両条件共にシフト長は 10msec)．辞書の基底はそれぞれの窓長に合わせ，Fourier 基底は窓長分の基底幅のみを用意し，Gabor 基底は 2^i の基底幅を持ち ($1 \leq i \leq s$, 窓長 25msec の場合 $s = 8$, 100msec の場合 $s = 10$)，基底が窓中を動く時間軸でのステップ幅は 64 サンプル刻みに設定した．MP での基底抽出処理は MPTK (Matching-Pursuit Toolkit) [2] を

A Study of Audio Feature Extraction Methods for Automatic Transformation of Environmental Sounds into Onomatopoeic Expression, *Nobuhide Yamakawa (Kyoto Univ.), Tetsuro Kitahara (Kwansei Gakuin Univ. / CrestMuse Project, CREST, JST), Toru Takahashi (Kyoto Univ.), Kazunori Komatani (Kyoto Univ.), Tetsuya Ogata (Kyoto Univ.), and Hiroshi G. Okuno (Kyoto Univ.)*

使い, それぞれの辞書で 64 基底を上記の音源から抽出し, 特徴量を得た.

隣接窓間の時間変化モデル化の効果を検証するために, HMM と GMM による同定性能を比較する. 両者共に混合数を 16, 遷移は Left-to-Right, HMM の状態数は 10 (中間層 8) 状態に設定した. 学習・識別処理には HTK (Hidden Markov Model Toolkit) [5] を使用した. クラス数は音源数と同じ 8 で, 評価は 10-fold cross validation で行った.

3.2 実験結果

窓内定常性有無 (Fourier or Gabor) と隣接窓間定常性有無 (HMM or GMM) それぞれの対ごとの同定率を, 基底抽出基底数を横軸にとって図示した. 図 1 は窓長が 25msec, 図 2 は窓長を 100msec に伸ばした場合の結果である. Fourier・Gabor 各基底の HMM での識別結果を音源別に並べたものを図 3(a) (窓長 = 25msec), 図 3(b) (窓長 = 100msec) に示す. 得られた結果を以下にまとめる.

1. 図 1 において, Fourier 基底では GMM を使った時, HMM よりも大きく同定率が下がる.
2. 図 2 において, Gabor 基底と HMM ペアの同定率が大きく向上し, Fourier 基底の両識別器での性能が下がっている.
3. 金属系衝突音の同定率が, 図 3(a) では Fourier 基底が優位だが, 図 3(b) ではその関係性が逆転している.
4. 複数窓を通じて定常的な音 (スプレーなど) は, 同定率が常に高い.

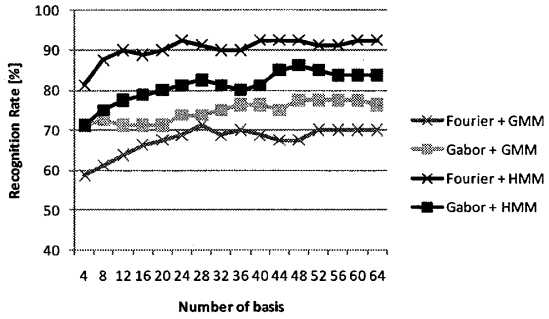


図 1 窓長 **25**msec での, 基底数ごとの基底と認識器の組み合わせ別同定結果

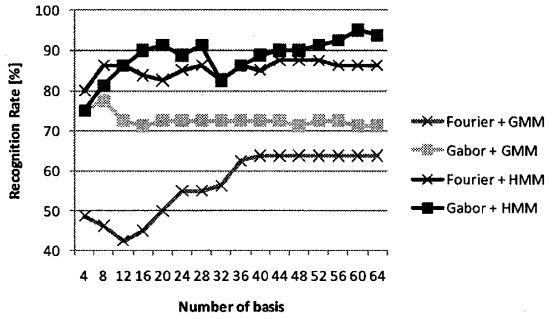
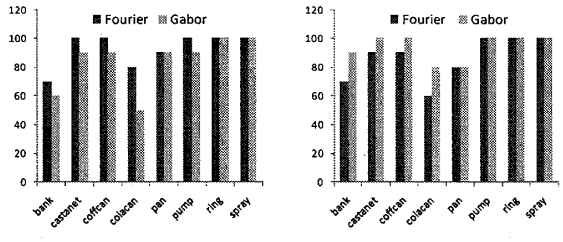


図 2 図 1 と同じ条件で窓長を **100**msec に変更した場合の結果



(a) 窓長: 25msec　　(b) 窓長: 100msec

図 3 Fourier/Gabor 各基底を用い HMM で識別させた場合の音源別結果

3.3 結果考察

- 結果 1 は, 窓長が短い場合に窓内の定常性がほぼ仮定できるが, 隣接窓間での時間変化が GMM ではモデル化できていないことに起因する.
- 結果 2 は, 窓長が 100msec では窓内の定常性が仮定できず, Fourier 基底が窓内スペクトラムの時間変化などを捉えられていないことを意味している.
- 窓 (100msec) 内の非定常性は, 実際に耳で音源中の 1 窓を聴いてみて, コーラ缶では減衰が有りスプレーでは持続的な音が聴取できたことからも確認できた.
- 結果 3 は, 窓長が長い時は窓内の定常性が仮定できないことを確認できる結果になっている. また Gabor 基底が図 1 で Fourier 基底より同定率が低いことの説明にもなっている.

以上で得られた知見により, 窓長をより長くしても Gabor 基底で信号の特徴抽出が行えるということがわかり, MFCC では時間変化検出のために短く設定されていた窓長を, より長くして音源同定性能の検討が可能になる.

4. 終わりに

Gabor 基底を用いて窓内の局所的な特徴を捉えることによって, Fourier 基底では同定の難しい分析窓内での時間変化が激しい (非定常) 信号でも, 上手く同定が行えることを確認した. また隣接窓間の時間変化を HMM でモデル化することで, 環境音の同定率が向上するという結論を得た.

今後はこの知見を擬音語認識に対しても適用させていく.

謝辞

本研究の一部は科研費の支援を受けた.

参考文献

[1] CHU, S. C. Environmental Sound Recognition With TimeFrequency Audio Features, *IEEE transactions on audio, speech, and language processing*, **17**, 6 (-08-01 2009), 1142.

[2] KRSTULOVIC, S. and GRIBONVAL, R. MPTK: Matching Pursuit made Tractable, Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'06), Vol. 3, Toulouse, France (May 2006).

[3] MALLAT, S. and ZHANG, Z. Matching pursuits with time-frequency dictionaries, *IEEE Transactions on Signal Processing*, **41**, 12 (1993), 3397–3415.

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

[5] YOUNG, S. and YOUNG, S. The HTK hidden Markov model toolkit: Design and philosophy, *Entropic Cambridge Research Laboratory, Ltd*, **2** (1994), 2–44.

[6] 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, **20** (20051101), 229–236.