ロボット聴覚のための Matching-Pursuit による環境音の分離音認識

○山川暢英† 高橋徹† 北原鉄朗‡ 尾形哲也† 奥乃博†
† 京都大学大学院情報学研究科
‡ 日本大学 文理学部

Separated Sound Recognition of Environmental Sounds using Matching-pursuit for Robot Audition

*Nobuhide Yamakawa† Toru Takahashi† Tetsuro Kitahara‡
Tetsuya Ogata† Hiroshi G. Okuno†
† Graduate School of Informatics, Kyoto University
‡College of Humanities and Sciences, Nihon University

Abstract—This paper presents the evaluation of a sound-source recognition method using time-frequency analysis and signal decomposition technique for environmental sounds with separation noise. Sound source separation is an essential technique for robot audition to recognize sounds from multiple sources. However, the technique has been developed particularly for speech recognition and the other sound categories such as environmental sounds have not extensively been involved. The separation process adds a distortion to a sound signal which result in deteriorated source recognition rate. To solve this problem, we investigated the validity of applying matching-pursuit with Gabor wavelets to separated signals. Experimental results show that, for sounds with flat spectrum, our method can retain high identification rate after a source separation while the rate decreases 25% when with MFCC.

Key Words: Robot audition, Sound source separation, Environmental sound recgnition, Matching-pursuit

1. はじめに

近年, コンピュータに様々な環境情報が含まれた混合音から有意な情報を引き出し, それに応じた行動をとらせるという目的から, 音環境理解 (Computational Auditory Scene Analysis) [1, 2] に基づいたロボット聴覚の研究が行われている [3, 4]. 音環境理解を実現するためには, (1) 音源方向を認識する **音源定位**, (2) 複数音源からの信号を分離する **音源分離**, (3) 分離された音を記号化する **音源認識** が必要であり, 個々の要素技術は独立した課題として研究されてきた. “音響処理の OpenCV” を目指して開発されているロボット聴覚ソフトウェア **HARK**[5] では, これらの技術を統合してシステム化している. 現状では主として音声に特化しており, 複数話者の発話内容を同時に認識するなどの音環境理解を実現している. しかし我々が日常生活で知覚する音には, 音声の他に音楽や環境音といった非音声音が存在し, それらを音声と別のものと認識し理解ができなければ, 本当の意味で音環境理解が実現したとは言えない. 特に環境音を理解することは, 異常音による危険察知や環境音イベントによる音源名学習など様々な利点が考えられる.

環境音は音声と音楽とは異なり, 多種多様な発音構造を持った音源を含むため, 繰り返し音, 定常性, 調波構造の有無なと音響的特徴が多岐に渡る. 各音カテゴリの代表的な音響的特性を表 1 にまとめた. これらの差異によって, 環境音の理解には, 音声/楽器音認識で用いられる技術とは異なった音響特徴量の設計や, 音源認識手法の開発が必要であると考えられる.

Table 1 音声, 楽器音, 環境音の音響的特性

Acoustical Characteristics	Voice	Music	Environmental Sounds
No. of Classes	No. of Phonemes	No. of tones	Undefined
Length of Window	Short (fixed)	Long (fixed)	Undefined
Length of Shift	Short (fixed)	Long (fixed)	Undefined
Bandwidth	Narrow	Relatively Narrow	Broad Narrow
Harmonics	Clear	Clear	Clear Unclear
Stationarity	Stationary	Stationary (except percussions)	Non-stationary Stationary
Repetitive Structure	Weak	Weak	Strong Weak
.	.	.	.
.	.	.	.

またロボット聴覚での使用を考えた場合, 現状の音源分離手法では分離音が歪むことを避けられない. 音声認識において最もよく使われている音響特徴量である Mel-frequency cepstrum coefficient (MFCC) では, 歪み成分が多ければそれだけ認識率も下がるため, “分離音歪みの影響を受けにくい” ことも特徴量設計の条件に加わる.

本稿では, 非定常性 (突発性) , 定常性, 複数発音などの特性を持った環境音及びその分離音を用意し,

1. 信号から非定常な特徴を抽出できる.
2. 分離歪み対する頑健性を持つ.

以上二つの条件に適した音響特徴抽出手法として時間周波数解析手法である Matching-pursuit (MP) [6] を検討する．本手法は時間周波数領域で局在化した信号，即ち非定常成分を検出し特徴量として扱えるという利点がある．また雑音への頑健性も報告されている [7]．

以下，第 2 章で音響信号の非定常性の定義と音源分離歪みの信号への影響を説明し，第 3 章で MP のアルゴリズムについて概説する．第 4 章では本手法による音源同定性能を従来法である MFCC と比較し有効性を調べる．

2. 音響信号の非定常性と分離歪み

本節では，環境音が持つ音響特性の例としてパワースペクトル変化の非定常性を紹介し，音源分離に Geometric Source Separation (GSS) [8] を使用した場合に発生する分離音のスペクトル歪の影響を概観する．

2·1 突発性環境音の音響特性

本稿では音響信号における非定常性を，“時間周波数領域で局在している成分の多さ” とする．定義に従えば，パワースペクトルの時間変化が大きく且つ細かい信号は非定常性が強いといえる．具体例として硬貨が擦れる時に生ずる “チャリチャリ” という音のスペクトログラムを以下に示す．

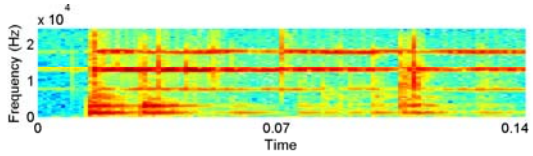

Fig.1 硬貨の摩擦音を細かい時間シフトで STFT した例: フレーム幅 = 8 msec, シフト幅 = 4 msec

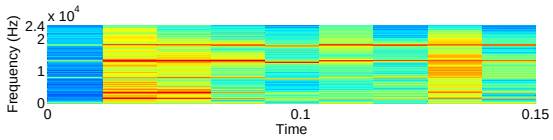

Fig.2 図 1 の信号を粗い時間シフトで STFT した例: フレーム幅 = 25 msec, シフト幅 = 10 msec

図 1 は信号を時間領域で細かく見た場合 (窓幅 = 8 msec, シフト幅 = 4 msec) を表している．13kHz と 18kHz 近辺に信号全体を通して高いパワーの信号が存在している．しかし硬貨同士が接触する度に，スペクトル全体に拡がる信号が生じ，高周波数成分 ($\geq$ 8kHz) が即座に減衰した後，その後最長で約 18 msec かけて低周波数成分 ($<$ 8kHz) が減衰している．

一方で，時間領域を粗く見た場合 (図 2)，図 1 では確認できたパワースペクトルの時間変化が，分析フレーム長が長いため消滅してしまっている．しかし周波数領域での解像度は増加しており，図 1 と比べ周波数軸上で局在化した成分を検出できている．

この様に STFT などを用いて時間領域で詳細な情報を得ようとすれば，周波数領域の情報が失われ，逆に周波数領域の解像度を上げた場合に，時間領域での情報が失われてしまう (不確定性原理)．時間周波数領域で両方の解像度を同時に増やせる手法としてウェーヴレット解析 [9] がある．本稿ではウェーヴレット解析を用いた音響特徴抽出手法として，Matching-pursuit を使用する．

2·2 音源分離による信号への影響

GSS はブラインド音源分離の一種であり，音源位置に関する幾何的拘束条件を設け，周波数領域で信号を分離するアルゴリズムである．従って，分離音源の認識で MFCC などのスペクトルベースの特徴量を使用した場合，非線形のスペクトル歪が全 MFCC に拡散してしまい，認識性能に悪影響を及ぼす．また GSS は残響成分を抑圧する作用も持つため，ここでも分離前後で特徴量の変化が生じる．

残響のある部屋の伝達関数が畳み込まれた電話の着信音 (約 2.8kHz で断続的に鳴るビープ音) を GSS を使用して分離した．図 3 と図 4 は，それぞれ分離前後の音源のスペクトログラムである．分離歪みと単独音データの S/N 比は約-4.3dB であり，スペクトログラムを比較すると，全体の背景ノイズと残響成分の減少が確認できる．

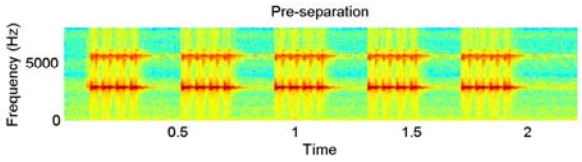

Fig.3 GSS による分離前の信号: 電話の着信音

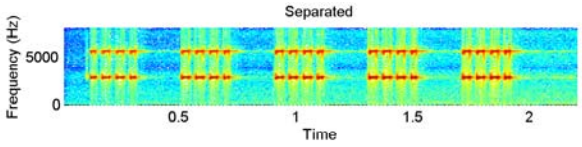

Fig.4 GSS による分離後の信号: 電話の着信音

3. Matching-pursuit と Gabor ウェーブレット基底による特徴抽出

本節では，MP のアルゴリズムと特徴量に用いる場合の処理について概説する．

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

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

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 $D = \{\phi_{\gamma_1}...\phi_{\gamma_{m'}}\}$ 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．

基底信号に式 (2) で表される離散 Gabor ウェーヴレット基底を使用した場合，それぞれの抽出基底に基底の時間幅 (s)，中心周波数 (ω)，時間位置 (u)，位相 (θ) の離

散値が保存されており, そこから必要な情報を取り出して特徴量とする.

$$g_{s,u,\omega,\theta}(n) = \frac{K}{\sqrt{s}} e^{\frac{-\pi(n-u)^2}{s^2}} cos(2\pi\omega(n-u)+\theta) \quad (2)$$

ここで K は $||g_{s,u,\omega,\theta}||^2 = 1$ となるような正規化項である.

準時間周波数解析手法である短時間フーリエ変換 (STFT) では, 分析窓の s と u の値を固定するため, フーリエ変換の不確定性原理により, 時間または周波数どちらかの軸上で局在化した信号しか検出できない. MFCC などの音響特徴量も STFT をベースにしており, 非定常な信号成分を抽出する目的に適した手法とは言えない. 一方, Gabor 基底は s を動かすことで時間方向の伸縮 (scale), u で時間軸方向の移動 (shift) を表現でき, 同時に周波数軸上の値を ω で調整できるので, 時間周波数領域で局在化した成分との相関がとれる. さらに m 個の信号成分をエネルギーの高い順番で抽出していくので, 高エネルギー成分が有意な特徴を持つと仮定した場合に, 信号分解のアルゴリズムともみなせる.

また MP における基底辞書はパラメータの解像度を自由に設定できるだけでなく, 異なる種類の基底信号を格納し信号解析に利用できる. その反面, MP で抽出した特徴による識別性能は基底辞書の記述内容に依存する.

本稿では, 比較対象とする MFCC に合わせ, MP でも 25 msec の分析窓を用い 10 msec 間隔のシフト幅で音響信号を解析する. Gabor 基底は $s = 2^p (1 \leq p \leq 8)$, $u = 16a (0 \leq a \leq 16)$, $\omega = Ki^{2.6} (1 \leq i \leq 32)$ を持つように設計した. また抽出処理には Matching-pursuit Toolkit (MPTK)[10] を使用した. 中心周波数と時間位置を特徴量とし, 抽出基底数× 2 の次元数を持った特徴ベクトルを実験に使用する.

4. 実験

8ch マイクアレイを持つロボットに搭載されたことを想定し, 8 入力の情報で分離した音源に対する認識率を, (1) 特徴量に MFCC を使用した場合 (音声認識の手法) と (2) MP と Gabor ウェーヴレットを使用した場合 (本手法) とで比較し, 本手法の非定常信号と分離歪みに対する有効性及び頑健性を検証する.

4·1 実験条件

4·1.1 使用音源

比較実験用の音源には, RWCP 実環境音・音響データベース [11] の非音声音ドライソースから, 衝突音系 (2 枚の木板, 金属板と金属棒, 5 種類のサイコロ, コップと木棒, 多数の粒と金属板, 拍手 (数回), コインと木板, 太鼓, 5 種類の本と紙) , 持続音系 (電話のビープ音, スプレー) の合計 11 クラスの音源を用いた. 音源は全てモノラル, 16bit/16kHz でサンプリングされている. それぞれ音源の物体や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.

4·1.2 音源分離

伝達関数には 8ch マイクアレイの搭載された川田工業の HRP-2 のものを使用した. 各マイク位置での伝達関数を使用音源に畳み込み, 残響のある部屋での録音を再現した分離前音源データ (計 8 個) を用意した. これらのデータから音源分離を GSS で行い, モノラルの分離後音源を生成した. 予備実験で今回の実験セット (単音の分離) では認識結果の音源到来方向への依存性が大きくないことを確認したため, 0° (正面) 方向から到来する音源のみを評価対象とした.

4·1.3 特徴抽出

- 分析フレーム: frame=25 msec, shift=10 msec
- MP: Gabor ウェーヴレット
  (抽出 48 基底 × 2 パラメータ = 96 次元)
- MFCC: 12 次元 MFCC+ΔMFCC+ΔPow=25 次元

4·1.4 識別器

識別は Left-to-Right, 3 状態 16 混合の Hidden-Markov Models (HMMs) で行った. 処理には Hidden-Markov-Model Toolkit[12] を使用した. 学習と識別は分離前後の音源で独立して行い, 10-fold-cross-validation で評価を行った.

4·2 実験 1: MFCC の分離/非分離音認識率比較

図 5 に特徴量に MFCC を用いた場合の認識結果を示す. 横軸はクラス名, 縦軸は認識率を表す. 濃い緑のグラフは分離前のデータを, 薄いグラフは分離後のデータを表している. “metal”, ”dice”, “cup”, “coins” などの金属との衝突音や音が複数回鳴る音源で認識率が 90%以下となり, 他の持続系の音や, 木などの柔らかい素材同士の衝突音のそれより低い. また”particles” は分離後に認識率が 25%下がっており, 逆に”phone-beep” では分離処理を行うことにより認識率が約 35%改善している. 全体の認識率は, 分離前のデータで 89.7%, 分離後のデータで 91.5%を示した.

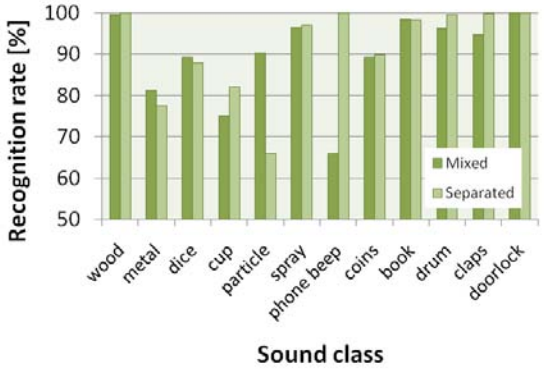

Fig.5 MFCC で特徴抽出した場合の非分離 (Mixed)/分離 (Separated) 音認識率

図 6 に分離前後で認識率の変動が多かった “particles” と”phone-beep” が, 認識率が下がった場合にどの音源と誤認識をしていたかを表したグラフを示す. 左グラフは音源分離後の”particles” が”spray” に 36.08%誤認識されていたことを表しており, 右グラフが音源分離前に”phone-beep” が”spray” に 46.15%誤認識されていたことを表している.

4·3 実験 2: MP の分離/非分離音認識率比較

図 5 に特徴量に MP/Gabor を用いた場合の認識結果を示す. 横軸はクラス名, 縦軸は認識率を表す. 濃い

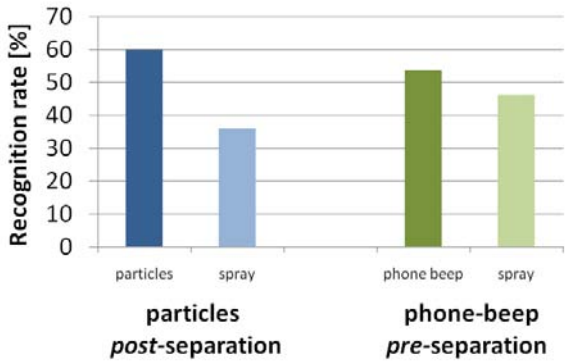

Fig.6 分離前の”phone-beep” (右グラフ) と分離後の”particles” (左グラフ) の認識結果において誤認識率が高かった音源名とその誤認識率

赤のグラフは分離前のデータを, 薄いグラフは分離後のデータを表している. “metal”, “dice”, “cup”, “coins” など, MFCC などでも認識率の低かった音が, MP ではさらに低い認識率を示した. 図 5 において分離前後で認識率が大きく変動した”particles” と”phone-beep” は, MP では分離前後両方で 90%以上の認識率を示し, 顕著な変動は見られなかった.

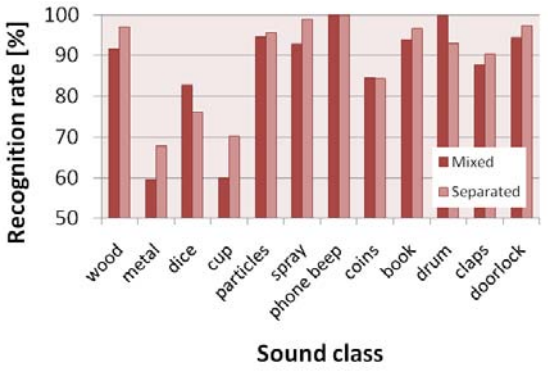

Fig.7 MP/Gabor で特徴抽出した場合の非分離/分離音認識率

4·4 考察 1: 非定常信号への MP/Gabor の効果

非定常性が強い信号に対しての認識率は, MFCC が MP を全体を通して上回った結果となったため, 本稿の実験条件では, 非定常環境音の認識タスクで MFCC に対する MP/Gabor の優位性は確認できなかった.

4·5 考察 2: 分離音への効果

“particles” は分解後で “spray” との誤認識率が増加していたことから, 分離歪が全 MFCC に拡がった結果, 元々平坦に近いスペクトル形状を持っていた両者の MFCC 距離が近づいたと考えられる. 実際に誤認識されていたファイル 10 個で MFCC 距離を測ったところ, 音源分離後に平均で 0.76 の距離の短縮が見られた.

一方 MP では分離歪や残響抑圧の影響が少なく, 分離後で約 1%認識率が向上した. これは元信号成分のエネルギーが分離歪のそれより大きいため, MP が抽出した基底が高エネルギーを持つ元信号の特徴を多く含み, 結果的に分離歪の影響が少ない特徴抽出ができているからだと考えられる. ここで抽出基底数を増やすにつれ, 歪み成分を多く特徴に含んでしまうことになり, 少なくすれば元信号の特徴を上手く表現できないため, 抽出基底数の設定には注意しなければならない.

5. 結論

本稿では, 非定常性と音源分離歪みを持った環境音音源を, MP と Gabor ウェーヴレットを使って認識する手法の有効性を検証した. 25 次元の MFCC が持つ認識性能と比較して信号の非定常性に対する優位性は確認できなかった. しかし MFCC では音源分離の影響を受け分離前後で認識率に大きい変動があるような音源でも, 本手法を用いることでその影響を少なくして特徴抽出が可能なことが明らかになった.

今後は, 音源の条件を変えて本手法の有効性を再度調査していく予定である. 例えば, 本稿では単音を分離し認識したが, より分離歪みが増える混合音に本手法を適用したり, 音源により多くの残響や白色雑音を加えるなどの条件で実験を行うことを検討している. また各条件で最適なウェーヴレット基底及び MP の抽出基底数も検討していくことも今後の課題である.

参考文献

[1] G.J. Brown and M. Cooke. Computational auditory scene analysis. *Computer speech and language*, Vol. 8, No. 4, pp. 297–336, 1994.

[2] D.F. Rosenthal and H.G. Okuno. *Computational auditory scene analysis*. L. Erlbaum Associates Inc., 1998.

[3] K. Nakadai, T. Lourens, H.G. Okuno, and H. Kitano. Active audition for humanoid. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 832–839, 2000.

[4] 奥乃博. ロボット聴覚の現状と展望. 日本ロボット学会誌, Vol. 28, No. 01, 2010.

[5] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and Implementation of Robot Audition System’HARK’Open Source Software for Listening to Three Simultaneous Speakers. *Advanced Robotics, 24*, Vol. 5, No. 6, pp. 739–761, 2010.

[6] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Process.*, Vol. 41, No. 12, pp. 3397–3415, 1993.

[7] S. Chu, Narayanan, S., and C.C.J. Kuo. Environmental sound recognition with timefrequency audio features. *IEEE Trans. Audio, Speech, Lang Process.*, Vol. 17, No. 6, p. 1142, 2009.

[8] L. C. Parra and C. V. Alvino. Geometric source separation: Mergin convolutive source separation with geometric beamforming. *IEEE Trans. Speech, Audio Process.*, Vol. 10, No. 6, pp. 352–362, 2002.

[9] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. on information theory*, Vol. 36, pp. 961–1005, 1990.

[10] Sacha Krstulovic and Rémi Gribonval. MPTK: Matching Pursuit made tractable. In *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP’06)*, Vol. 3, pp. III–496 – III–499, Toulouse, France, May 2006.

[11] 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.

[12] S. Young and S. Young. The HTK hidden Markov model toolkit: Design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, Vol. 2, pp. 2–44, 1994.