

# ロボット聴覚のための Matching Pursuit による複数環境音の同定

山川 暢英<sup>†</sup>高橋 徹<sup>†</sup>北原 鉄朗<sup>‡</sup>尾形 哲也<sup>†</sup>奥乃 博<sup>†</sup><sup>†</sup> 京都大学大学院 情報学研究科 知能情報学専攻<sup>‡</sup> 日本大学 文理学部 情報システム解析学科

## 1. はじめに

近年、大きく進歩してきたロボット制御技術に呼応して、画像だけでなく聴覚を含めた実世界での認知機構の研究が盛んになっている。ロボットの耳で聞き分ける“ロボット聴覚”は、オープンソースソフトウェアの HARK [1] が公開され、徐々にロボットに搭載され始めている。これまでにロボット聴覚の対象になっているのは、音声の主であり、1つの音声を聞き分けるカクテルパーティ効果ロボットや、複数の音声を同時に聞き分ける聖徳太子ロボットなどのデモが行われている。

実世界に配備されるロボットが聞く音は、複数の音声だけでなく、音楽あるいは環境音、そしてそれらの混合音である。実環境では音声以外の音も多くの情報を含んでおり、音を通じて環境認識する音環境理解が不可欠である。非音声は周囲環境を反映したサインとして機能する。例えば、人間同士の会話の最中に時計アラームが鳴ると、誰かがその音源と場所を指示し、別の人がアラームを止め、会話が再開される、という流れがありえる。こうした対話にロボットを参加させるためには、非音声、特に環境音認識機能が必要となる。

従来の環境音認識研究はロボット聴覚とは直接的関係が薄く、動物の鳴き声の音響特徴から心理状態を推定する研究 [2] や、モバイル機器に現在地の周囲環境を同定させる目的で、街の雑踏やカジノなどの背景音同定研究 [3, 4]、異常音を他の環境音と識別するものが主流である。特に家庭用セキュリティロボットなどの応用を見据えた異音検知の研究も進められている [5]。しかしいずれも実環境でロボットに入力される信号として単一音源が想定されており、複数の方向性音源と背景雑音との混合音であるという条件を考慮していない。

本稿では、雑音環境下で同時に別方向から発生する複数の環境音を同定する実験を通じて、音源定位と分離で音の聞き分けをするロボット聴覚の性能と、音源同定に適した音響特徴量を報告する。

## 2. ロボット聴覚における環境音認識

本稿では環境音を2種類に大別する。1つ目は、短時間で変動の大きい、チャイムやドアノック音など特定の意味を持つ音で、ここでは“音イベント”と称する。方向性を持ち、複数個の音イベントが同時に発生し得る。2つ目は、長時間に渡り比較的変動の少ない音で一般的に“背景雑音”とされ、エアコンの動作音や居酒屋の喧噪などが該当する。エアコンの音が通常存在感以上の意味を持たない雑音と見なされるのに対し、居酒屋の喧噪はロボットの所在環境を音から判断する材料となる。背景雑音はロボット自身の動作音(例、ファンノイズ)も含む。これら2種類の環境音は常に同時に発生し、ロボットはこれらを分離し、すべてを認識しなければならない。

ロボット聴覚による音源分離では、空間的スパースネスを前提に分離するので、指向性を持つ音イベントは分離されるが、無指向の背景雑音は分離後の信号に残留するため、認識系への入力信号は単一の音イベントと背景

雑音の混合音となる。環境音認識では、背景雑音の影響に頑健な特徴量の設計が必要となる。人間の音声認識で通常使用される mel-frequency cepstral coefficients (MFCC) は、背景雑音が多い時や、信号スペクトルの時間変化が速い(非定常な)音源に対して影響を受けやすいので、音イベントの認識に対し適切な特徴量ではない。背景音と異なり、音イベントは:

1. エネルギーが時間周波数領域で局在している、
2. エネルギーが大きく波形上でも突出している、

という特徴を持つ。

即ち時間周波数領域でエネルギーの突出した成分だけを抽出する手法が望ましい。

### 2.1 特徴抽出

本稿では、上記の条件を満たす音響特徴抽出アルゴリズムとして、matching-pursuit (MP) を使用する。MP は所与の信号  $s$  から時間周波数領域の成分  $\phi_\gamma$  をエネルギーの高い順番に  $m$  個抽出可能なアルゴリズムであり、その数を適切に設定すれば、時間周波数領域でエネルギーの弱い背景雑音を無視し、突出した成分だけを抽出できる。従って音響特徴量に雑音頑健性を期待できる。

抽出された成分  $\phi_{\gamma_1} \dots \phi_{\gamma_m}$  はそれぞれ任意の辞書  $D$  に記述された基底  $\{\phi_{\gamma_1} \dots \phi_{\gamma_{m'}}\} (m' \geq m)$  で表現され、式 (1) のように線形結合で信号を近似できる。

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

ここで  $R^{(m)}$  は残差信号を表す。抽出基底の選択処理は:

1. まず  $D$  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  $\phi_{\gamma_1}$  としてその相関係数  $\alpha_{\gamma_1}$  と共に  $s$  から抽出し、
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、 $\alpha_{\gamma_2} \phi_{\gamma_2}$  を得て、
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。

抽出された基底は、基底幅、周波数、時間位置などのパラメータを持つため、これらを特徴量ベクトルとして扱える。抽出処理の計算量は  $m$  に比例するため、抽出効率の良い辞書を設計し少ない  $m$  で有意な特徴を得ることが理想である。アルゴリズムの詳細は [6] を参照されたい。

#### 2.1.1 背景雑音に対する頑健性

例として雑音を含む音響信号から高エネルギー成分を MP を用いて抽出する過程を図 1 に示す。信号には左図に示す約 3500Hz で持続的に強いエネルギーを持つチャイムの音を使用した。その信号に背景雑音の典型として白色雑音 (SNR=0dB) を加えた信号 (中図)、そこから 48 個の Gabor 基底を MP で抽出した再合成信号 (右図) をスペクトログラムで示す。中図では 3500Hz 付近の山を除き、全てのフレームでスペクトルが雑音に埋もれてしまっている。こういった信号からフーリエ変換を使用して特徴抽出を行った場合、雑音成分を特徴に含んでしまい、特徴ベクトルは原信号のそれとは大きく異なってしまふ。従って音響信号の認識手法で一般的な MFCC を

Multiple Sound-sources Identification using Matching Pursuit for Robot Audition, Nobuhide Yamakawa (Kyoto Univ.), Toru Takahashi, Tetsuro Kitahara (Nihon Univ.), Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)

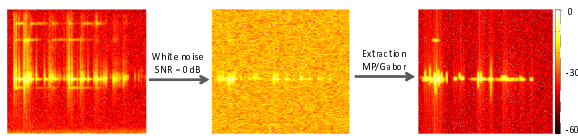


図1 Signal extraction by matching-pursuit with the Gabor wavelet

表1 Experimental setup (sound classes, audio features, classifiers)

	MFCC	MP/Gabor
Audio features	MFCC(12) + ΔMFCC(12) + Δpower(1) Window width = 25 msec Shift rate = 10 msec	Frame
		Window-width = same as MFCC Base-width = 2-256 samples Time-shift resolution = 16 samples # of extracted atoms = 48
Classifiers	GMMs (4-mixes) or HMMs (6 hidden states)	
Sound classes	one at 0 degree (in front) and the other at 45 degree off	
	chime, handclap, clock-alarm, coin-clinks, glass-cups, doorknob, metal-plate, phone-beep, bell-ringing, metal-bin	
Learning set	Original signal (clean) + Noised signals (SNR = 0dB) + Noised separated signals = 900 samples	
Test set	Separated signals = 100 samples	

含むフーリエ変換を核にした手法では背景雑音の影響で認識性能が低下する。一方で右図のMPによる再合成信号では、背景雑音を排除しながら3500Hz周辺の成分だけを抽出している。このように、MPを使用することで、信号中の突出した特徴を多量の背景雑音の中から抽出できる。

### 3. 同定性能の比較実験

実環境においてロボットは種々の音源や背景雑音に囲まれており、ロボット聴覚ではそれぞれを個別に認識するために音源定位及び分離等の信号処理を行う。本実験の目的は、そうした信号処理の影響を含んだ環境音信号に対するMPの音響特徴抽出の性能をMFCCと比較検討し確認することである。

#### 3.1 実験条件

比較実験用の音イベントには、RWCP実環境音・音響データベース[7]の非音声源ドライソースから、衝突音などの減衰の速い音8種と、アラーム音などの減衰しない音2種の合計10種を使用した。音源は全て16bit/16kHzでサンプリングされており、信号長は平均0.7秒である。それぞれ同じ音源を無響室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。実験では、音源分離処理と音イベントに方向性を持たせるために、8チャンネルマイクアレイで実測したロボットの頭部伝達関数(HRTF)を使用した。HRTFを畳み込み白色雑音(SNR=0dB)を加算することによって、異なる到来方向を持った2つの音イベントと背景雑音の混合音を生成する。

次に、MUSIC法による音源定位で得た各音イベントの幾何情報を使い、幾何制約付き高次無相関化音源分離法(GHDSS)を用いて音イベントを音源分離する。これらのアルゴリズムにはHARKのモジュールとして実装されているものを利用した。表1に音響特徴量、音源、学習/評価条件の詳細を示す。

MPの基底にはGaborウェーブレットを使用し、抽出数とHMMの状態数は過去の実験[8]から最適な性能を示す組み合わせを選んだ。なお、MPは信号全体ではなく、MFCCと同様にフレームごとに適用される。以上の条件で分離信号の音源同定を行い、各識別器の同定正解率を10交差確認法により算出した。

#### 3.2 実験結果

結果に音源到来方向への顕著な依存性が見られなかったため、45°方向から到来する音の同定率のみを図2に

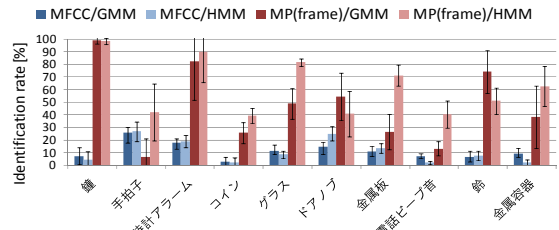


図2 到来方向 45°音源の同定正解率の比較: MFCC/GMM, MFCC/HMM, MP(frame)/GMM and MP(frame)/HMM

示す。横軸にはクラス名が列挙してあり、縦軸は同定率を表している。全体の傾向として、雑音環境下ではMPがMFCCの同定性能を大きく上回っており、ドアノブと鈴以外の音源でHMMで識別したものがGMMよりも同定率が高かった。同定率が落ちた音源に共通する特性は、7kHz以上にパワーのピークが存在することである。分離音源はGSSによるローパスフィルタ効果で高周波数領域の特徴が失われる。これにより分離前後の信号特徴に差異が生じてしまい、HMMによる識別では別の音源として判別されてしまう。一方でGMMは特徴ベクタの時間遷移を考慮しないため、信号全体で最も頻繁に入力された特徴が支配的となる。ドアノブや鈴といった衝突音は高周波数より低周波数成分が長く持続する減衰パターンを持つので、結果としてGMMは分離前後で共通して強い低周波数の特徴を重点的に学習し、HMMより高い同定性能を示す。

また大別して90%近くの同定率を示す音源と、50%以下の低い同定性能を見せる音源がある。同定率の良いクラスは調波構造が明瞭であり、クラス全体で周波数特徴の分散が小さいが、手拍子、コイン、電話ビープ音といった同定率の悪いクラスは、調波構造が不明瞭か音高情報がクラス全体で大きく変化する傾向にあった。従って、MPで抽出できる周波数特徴をこれらの音源に適用することは適切ではない。

### 4. 終わりに

実環境で想定されるような背景雑音環境下でのロボット聴覚による複数環境音の同定実験を行った。特徴量としてMPとGaborウェーブレットを用いることで背景雑音に対する頑健性を示したが、分離処理の影響や音源自体の特性により、識別手法や特徴量を適切に設定する必要性を確認した。今後は種々の音響特性に適した音源同定手法の設計と考察を進めていく。

謝辞 本研究は科研費(S)(A), GCOEの支援を受けた。

### 参考文献

- [1] 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.
- [2] Y. Ikeda and Y. Ishii. Recognition of two psychological conditions of a single cow by her voice. *Computers and Electronics in Agriculture*, Vol. 62, No. 1, pp. 67-72, 2008.
- [3] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE TASLP*, Vol. 14, No. 1, pp. 321-329, 2005.
- [4] S. Chu, S. Narayanan, and C.C.J. Kuo. Environmental sound recognition with timefrequency audio features. *IEEE TASL*, Vol. 17, No. 6, p. 1142, 2009.
- [5] S. Ntalampiras, I. Potamitis, and N. Fakotakis. Sound classification based on temporal feature integration. *ISCCSP-2010*, pp. 1-4, 2010.
- [6] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE TSP*, Vol. 41, No. 12, pp. 3397-3415, 1993.
- [7] 実環境音声・音響データベース. Rwcpsound scene database in real acoustical environments. <http://tosa.mri.co.jp/sounddb/index.htm>.
- [8] N. Yamakawa, T. Kitahara, T. Takahashi, K. Komatani, T. Ogata, and H.G. Okuno. Effects of modelling within-and between-frame temporal variations in power spectra on non-verbal sound recognition. *INTERSPEECH-2010*, pp. 2342-2345, 2010.