あらゆる音の検出・識別を目指して –音響イベント検出研究の現在と未来– *

○大石康智 (日本電信電話株式会社 NTT コミュニケーション科学基礎研究所)

1　はじめに

膨大な音や映像のメディアデータが身の回りにあふれる中，これらのデータを自在に解析し，例えば，ライフログや高齢者の見守りシステム，ハウスセキュリティシステムなど，人間の行動理解を目指した様々な応用を実現するためには，付随するテキストデータに頼るだけではなく，それぞれの中身を表す情報を，音や映像自体から自動的に引き出す技術が必要不可欠である。特に，音に含まれる情報は人間が置かれた環境（シーン）を理解する上で重要な手掛かりとなる。しかし，これまで我々が扱ってきた音は，主に音声，もしくは音楽に限定されており，その限られた世界での応用について議論が行われてきた。

これに対して，音シーン理解の研究ではその囲いを取り払い，あらゆる音を対象とするため，大幅な音アプリケーションエリアの拡大が見込まれる。それを裏付けるように，ここ数年で実世界に存在する様々な音を検出，識別する音響イベント検出（本稿では音シーン理解の研究の一つの課題と位置づける）が活発化しており，様々な成果が報告されつつある。Fig. 1 のグラフは，国際会議 ICASSP における，音シーン理解の研究の発表件数を示しており，10 年間で発表件数が徐々に増えていることが分かる（ここでは,「audio」「event」「sound」「scene」「detection」「classification」をキーワードとしてタイトルを検索し，筆者が内容を精査して音シーン理解の研究であると主観的に判断した件数を示す)。2013 年は音響イベント検出に関するスペシャルセッションが開催されたため，特に発表件数が多い。同時に Fig. 1 の下部では，10 年間に開催された競争型ワークショップを示す [1–5]。定期的に開催される競争型ワークショップが，研究の活発化と発表件数の増加を後押ししていると考えられる。

本稿では，このような情勢を鑑みて，音シーン理解に関する多くの研究を，ある 2 つの観点から眺めることによって，10 年間に渡って取り組まれてきた課題を整理する。次に，上記の競争型ワークショップが，どの課題に取り組むために開催されたかを調査し，現状の技術では解くことが難しい課題を明らかにする。最後に，ここ数年の間に洗練されてきた機械学習（深層学習やベイズ学習）を駆使して，この難しい課題に取り組む研究事例を紹介し，今後の研究の方向性を議論する。

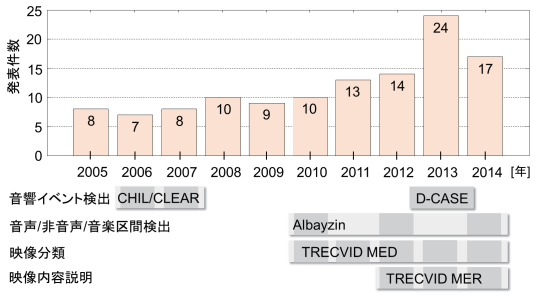

Fig. 1　国際会議 ICASSP における音シーン理解の研究発表件数と競争型ワークショップの開催状況

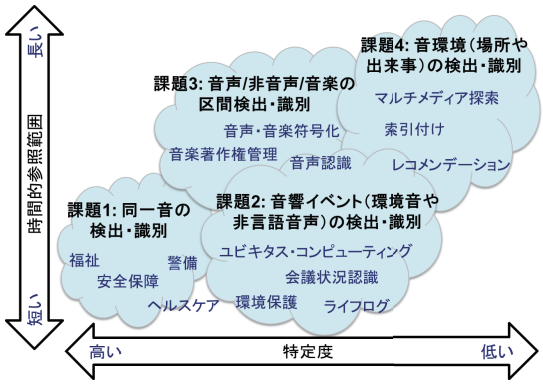

Fig. 2　特定度と時間的参照範囲の 2 軸で整理した音シーン理解の研究課題

2　音によるシーン理解の研究における課題

音シーン理解の研究課題を，Fig. 2 に示すような，「特定度」と「時間的参照範囲」の 2 軸で整理する。ここで,「特定度が高い」とはデータベースの音と全く同一の音を観測信号から検出して識別することを意味する。「特定度が低い」とはデータベースの音と何らかの観点で同一の音を観測信号から検出して識別する。一方,「参照範囲が短い」とは，短いフレーム単位で照合しながら，音を検出して識別する。「参照範囲が長い」とは，ある程度長い，信号全体の統計的な特徴に基づいて，音を検出して識別する。

この 2 つの軸は音楽検索における課題を分類するために用いられており [6]，その類推で音シーン理解の研究を整理すると，同一音の検出・識別，音響イベント（環境音や非言語音声）の検出・識別，音声/非音声/音楽区間の検出・識別，音環境（場所や出来事）の検出・識別という 4 つの課題に分類できる。各々の

---

* Toward detection and discrimination of all sounds –present and future of audio event detection–. by OHISHI, Yasunori (NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation)

課題について，目的とその応用を概説する。

2.1 課題 1: 同一音の検出・識別

データベースに集められた音と全く同一の音を観測信号から検出して識別する。警報ブザーや電化製品の音，信号機の音，銃声のような，とても短い単発音を正確に検出することが求められており，高齢者や障がい者への福祉用具，警備や安全保障などへの応用が期待されている [7–9]。雑音や信号歪みに頑健な音響指紋技術や，最近ではスペクトログラムを画像処理することによって得られる特徴を利用して照合する手法が増えている [10, 11]。他の課題に比べて研究の数は少ないが社会的ニーズは大きいため，今後活発に研究されることを期待する。

2.2 課題 2: 音響イベントの検出・識別

環境音や非言語音声からなる「音響イベント」をあらかじめ定義し，その音響的特徴の観点で同一の音を観測信号から検出して識別する。音源や収録環境，話者性の違いを含めて同じ音響イベントを検出するため，2.1 節に比べて特定度は低くなるが，単発音を対象とするため時間的参照範囲は短い。ライフログや会議状況認識，環境保護などへの応用を想定する。基本的にはメル周波数ケプストラム係数（MFCC）等を特徴量とし，隠れマルコフモデル（HMM）を利用して，時間区間を特定する [12–16]。最近では，非負値行列因子分解（NMF）のようなスパース信号解析によって得られる基底も音響特徴量に用いられる [17–19]。データベースを独自に収集する研究機関も多いが，一般的には，RWCP 実環境音声・音響データベース [20] における 105 種類の環境音や AMI Meeting Corpus[21] における笑い声のような非言語音声，60 時間に及ぶ BBC Sound Effects[22] が評価に利用される。

2.3 課題 3: 音声/非音声/音楽区間の検出・識別

音声信号と音声以外の信号（非音声信号）が含まれる観測信号から，音声信号が存在する時間区間を検出する音声区間検出（VAD）の研究は長い歴史を持ち，これまでに様々な手法が提案されている [23]。この課題は，特定度が音響イベント検出と同程度で，時間的参照範囲が長いと位置づけられる。VAD は，音声符号化や音声認識等の様々な音声技術に応用されており，ITU-T，及び ETSI より標準化された技術が幅広く利用されている [24, 25]。また，近年では音声と非音声の識別のみではなく音楽信号の識別を行う，汎用信号区間検出技術（GSAD）も ITU-T より標準化されている [26]。VAD の評価用データベースとして，CENSREC-1-C[27] が公開されており，GSAD においても評価用データが公開されている [28]。

2.4 課題 4: 音環境の検出・識別

「オフィス」や「地下鉄」,「ドッグショー」や「パレード」のように，場所や出来事をあらかじめ定義し，その音響的特徴の観点で同一の音環境を観測信号から検出して識別する。同じ場所や出来事でも個々の構造は極めて多様であるため，特定度は低いと言える。そのため，比較的長い音響信号における大域的な特徴を参照して識別する必要がある。動画共有サービスに投稿された大量の映像クリップに索引付けを行い，マルチメディア探索や推薦への応用を見据えている。基本的な手法は Bag of Words である [29–32]。MFCC をはじめとする様々な特徴量を抽出し，あらかじめベクトル量子化により作成されたコードブックを用いて，そのヒストグラムを作成する。そして，それを入力としたサポートベクターマシン (SVM) で識別する。評価用データベースとして，Columbia Consumer Video Dataset[33] が公開されている。

3 音によるシーン理解の研究を促進させる競争型ワークショップ

ここ 10 年間に開催された音シーン理解の研究に関連する競争型ワークショップが，前節のどの課題に取り組んでいるかを説明する。

3.1 CHIL/CLEAR [1]

CLEAR（CLassification of Events, Activities and Relationships) は，2006 年と 2007 年に VACE と CHIL の協賛の下で開催された競争型ワークショップである。タスクの一つとして，会議室におけるセミナーの様子を録音した音響信号から,「ドアノック音」,「足音」,「椅子の移動音」,「電話の呼び出し音」,「拍手音」,「笑い声」などの 12 個の音響イベント検出が行われた。Fig. 2 の課題 2 に取り組んでいる。6 チームが参加し，最も高い性能を得たチームは，MFCC をはじめとする様々な特徴量を AdaBoost によって選別し，HMM を利用して時間区間を特定した [34]。それでも全体的に性能は低く，原因は音響イベントが複数話者による音声と重なった状況にあったためである。多チャンネル信号を利用することが今後の方向性として挙げられた。あらかじめ切り出された 12 個の音響イベントの識別や,「空港」や「公園」など 9 つの音環境の識別は比較的簡単なタスクであったため，CLEAR 2007 では廃止された。

3.2 Albayzin [2]

Albayzin は，スペインの大学・研究機関によって，2006 年から隔年開催される音声言語処理の競争型ワークショップである。Albayzin 2010 以降，およそ 87 時

間分の TV ニュース番組の音響信号から,「音声」,「音楽」,「背景に雑音が重畳する音声」,「背景に音楽が重畳する音声」,「その他」からなる 5 つのクラスを区間検出して識別するタスクが行われている。Fig. 2 の課題 3 に取り組んでいる。MFCC やクロマ特徴量などの 1 秒程度の統計量を特徴量に加えること，HMM を使いながら 5 つのクラスを階層的に識別することの有効性が報告されている [2]。

3.3 TRECVID MED [3]

NIST が主催する TRECVID MED（Media Event Detection）は，映像クリップから音情報を含めた「イベント」を検出して識別するタスクであり，毎年 20 チームほどの研究機関が参加する。イベントとは，ある特定の時間・場所における，人間と人間や，人間から事物への行動や出来事を指す。「タイヤを交換している」,「誕生日を祝っている」,「岩山を登っている」などがその例である。Fig. 2 の課題 4 に取り組んでいる。2013 年は 5840 時間の映像データに対し，30 個のイベントが定義された。MFCC を特徴量，話者認識で開発されたガウス混合モデル（GMM）の Supervector と SVM の組み合わせを識別器とし，GMM の木構造探索を高速化に用いるといった，音声分野で開発された技術が性能向上に寄与することが報告されている [35]。最近では，意味インデキシングや音声認識，OCR によって得られた情報を「中間表現」とし，それらを入力とした検出器を従来法と組み合わせるアプローチが試みられている。

3.4 TRECVID MER [4]

TRECVID MED の参加者を対象に，なぜイベントが検出されたか証拠を列挙して「説明する」，MER（Media Event Recounting）が 2012 年から開催されている。映像クリップにおける証拠の場所と時刻，およびその説明文を，XML 形式で書き起こす。これは，MED の性能解析・向上とともに，検索インタフェースの利便性向上を目的とする。証拠が音響イベントにも相当することから，Fig. 2 の課題 2 に取り組んでいると言える。2013 年は 10 チームが参加しており，人手による判定で，60%程度のイベントの「説明」が可能であることが報告されている [4]。

3.5 D-CASE [5]

D-CASE（Detection and Classification of Acoustic Scenes and Events）は，IEEE Signal Processing Society による後援の下，音環境識別と音響イベント検出を行った競争型ワークショップである。Fig. 2 の課題 2 と 4 に取り組んでいる。音環境識別では，屋内や屋外からなる 10 個の場所が定義された。11 チームが参加し，最も高い性能を得たチームは，再帰定量化解析を用いて MFCC 系列の動特性を特徴抽出し，SVM を用いて識別した [36]。この性能が人間による識別能力と同等であったことが興味深い。一方，音響イベント検出では，オフィス環境で観測される 16 個の音響イベントが定義された。評価では，実際の様々なオフィス環境で収録された 1 分程度の音響信号を利用した。また，時間的に重なった音響イベントの検出性能を検証するために，手動で合成した音響信号も評価した。7 チームが参加し，高い性能を得たチームは，MFCC や Gabor フィルタバンク出力値を特徴量とし，HMM を検出器とした [37]。全体的にイベント検出率の性能は低く，特に時間的に重なった音響イベントをほとんど検出できないことが明らかとなった。

4 機械学習を用いた音響イベント特徴抽出

5 つの競争型ワークショップを踏まえると，特に難しい課題は，観測信号から比較的短い音響イベントの時間区間を検出して識別することである。ここでは，音響イベントが，音声や他の音響イベントと時間的に重なっている状況を対象とする。これは，TRECVID MER のような，場所や出来事を識別した上で，その内容を事細かに説明するためにも重要な課題である。

筆者は，この課題に取り組む一つの重要な点が音響イベントの特徴抽出であると考える。本節では，SVM のような識別器だけでなく，特徴抽出器を内包する機械学習法を利用した，2 つの音響イベント検出手法を紹介し，今後の研究の方向性を議論する。

4.1 深層学習に基づく音響イベント特徴抽出

深層学習を利用した音声認識器がこれまでの最先端技術を結集した認識器の精度をさらに超えたこともあり，音環境識別や音響イベント検出に深層学習，特にディープニューラルネットワーク（DNN）が導入されはじめた [38, 39]。Espi らは，Fig. 3 のように，制約付きボルツマンマシン（RBM）に基づく自己符号化器によって事前学習された隠れ層を積み重ねて，多層の階層ネットワークを構築し，最終層の出力を使った識別ネットワークを追加して，全体として教師あり学習をさせて，音響イベント検出を行った [39]。事前に特徴抽出された MFCC やフィルタバンク出力値よりも，およそ 200ms 区間のスペクトルを並べたスペクトルパッチを入力層に使うことによって，検出精度が向上した。教師なし学習された隠れ層が，時間的に重なった音響イベントの詳細な特徴抽出の役割を果たしている。特徴抽出器と識別器の両方の機能を備えた DNN が，音響イベント検出の精度改善と新たな応用開拓に貢献することを期待する。

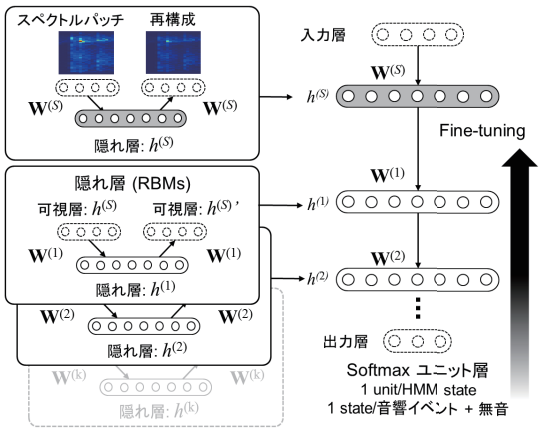

Fig. 3　深層学習に基づく音響イベントの検出・識別

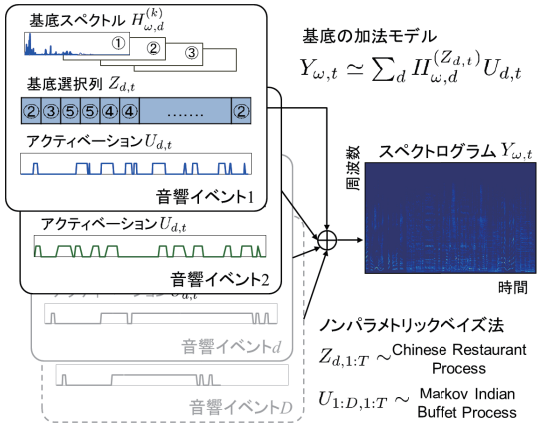

Fig. 4　ベイズ学習に基づく音響イベントの検出・識別

4.2　ベイズ学習に基づく音響イベント特徴抽出

音のスパース性や連続性などを事前知識として導入できる，ベイズ学習に基づく音響イベント検出も検討されている [40, 41]。さらに，音響イベントの特徴表現や音響イベントの個数に関するモデル選択問題を回避できるノンパラメトリックベイズ法の導入も検討される [42, 43]。これらの手法は事前知識や素朴な制約の下で，観測信号から音響イベントの音響的特徴を教師なしで学習できることを特長とする。Ohishi らは，Fig. 4 のように NMF を土台とし，音のスパース性と連続性の仮定の下で，観測信号のみから音響イベントの時間区間を検出する手法を提案した [42]。基底スペクトルと音響イベントの個数は，ノンパラメトリックベイズ法によって自律的に学習された。GMM に基づくベースライン法に比べて検出精度を上回った。検出対象の音響イベントの学習データが十分に与えられない状況において，過学習することなく，音響特徴抽出できる有望な手法となることを期待する。

5　おわりに

あらゆる音を対象とした音シーン理解の研究は，福祉や警備，マルチメディア探索など様々な応用を見据えて，今後も急速に進展していくことが予想される。本稿では，音シーン理解の研究を 4 つの課題に分類し，これらに関連する競争型ワークショップの取り組みを調査した。その結果，比較的短い音響イベントの検出・識別がまだまだ技術的に難しい課題であることが分かった。最後に，音響イベントの特徴抽出を内包する機械学習法を駆使して，この課題に取り組む研究事例を紹介した。これらの研究事例の課題は，高い検出精度を保ちながら，計算コストを抑え，データベースの規模に対する拡張性を高めることである。

謝辞　執筆にあたり，NTT 研究所 井本桂右氏，植松尚氏，大室仲氏，柏野邦夫氏，中谷智広氏， 藤本雅清氏，Miquel Espi 氏から有益なコメントをいただいた。

参考文献

[1] R. Stiefelhagen *et al.*, Multimodal Technologies for Perception of Humans, vol. 4625, pp. 3-34, 2008.
[2] T. Butko *et al.*, EURASIP J. Audio, Speech and Music Processing, 2011.
[3] http://www.nist.gov/itl/iad/mig/med.cfm
[4] http://www.nist.gov/itl/iad/mig/mer.cfm
[5] D. Giannoulis *et al.*, in *Proc. WASPAA 2013.*
[6] P. Grosche *et al.*, Multimodal Music Processing, pp. 157–174, 2012.
[7] A. Harma *et al.*, in *Proc. ICME 2005.*
[8] A. Pikrakis *et al.*, in *Proc. ICASSP 2008.*
[9] Y. T. Peng *et al.*, in *Proc. ICME 2009.*
[10] J. P. Ogle *et al.*, in *Proc. ICASSP 2007.*
[11] X. Lu *et al.*, in *Proc. ICASSP 2014.*
[12] S. Petridis *et al.*, in *Proc. ICASSP 2010.*
[13] F. Weninger *et al.*, in *Proc. ICASSP 2011.*
[14] S. Chaudhuri *et al.*, in *Proc. ICASSP 2013.*
[15] R. Chakraborty *et al.*, in *Proc. ICASSP 2013.*
[16] T. Heittola *et al.*, in *Proc. ICASSP 2013.*
[17] C. V. Cotton *et al.*, in *Proc. WASPAA 2011.*
[18] T. Heittola *et al.*, in *Proc. CHiME workshop 2011.*
[19] J.F. Gemmeke *et al.*, in *Proc. WASPAA 2013.*
[20] S. Nakamura *et al.*, in *Proc. ICME 2002.*
[21] I. McCowan *et al.*, in *Proc. Measuring Behavior 2005.*
[22] http://www.sound-ideas.com/bbc.html
[23] 藤本, 電子情報通信学会誌, vol. 95, no. 8, pp. 754-758, 2012.
[24] ETSI TS 101 707 v.7.5.0, 2000.
[25] ETSI ES 202 050 v.1.1.4, 2006.
[26] ITU-T Recommendation G.720.1, 2010.
[27] N. Kitaoka *et al.*, Acoust. Sci. & Tech., vol. 30, no. 5, pp. 363-371, Sept. 2009.
[28] http://www.itu.int/net/ITU-T/sigdb/speaudio/Gseries.htm#G.720.1
[29] S. Pancoast *et al.*, in *Proc. ICASSP 2013.*
[30] Z. Huang *et al.*, in *Proc. ICASSP 2014.*
[31] Y. Wang *et al.*, in *Proc. ICASSP 2014.*
[32] E. Amid *et al.*, in *Proc. ICASSP 2014.*
[33] Y. Jiang *et al.*, in *Proc. ICMR 2011.*
[34] X. Zhou *et al.*, Multimodal Technologies for Perception of Humans, vol. 4625, pp. 345-353, 2008.
[35] 篠田, 音講論集, 1-4-10, pp. 531–532, 2014.
[36] G. Roma *et al.*, in *Proc. WASPAA 2013.*
[37] J. Schroder *et al.*, in *Proc. WASPAA 2013.*
[38] Z. Kons *et al.*, in *Proc. INTERSPEECH 2013.*
[39] M. Espi *et al.*, in *Proc. HSCMA 2014.*
[40] A. Mesaros *et al.*, in *Proc. EUSIPCO 2011.*
[41] K. Imoto *et al.*, in *Proc. MLSP 2013.*
[42] Y. Ohishi *et al.*, in *Proc. ICASSP 2013.*
[43] Y. Sasaki *et al.*, in *Proc. WIAMIS 2013.*