情報科学類専門語学Aレポート
執筆上の注意について

ソフトウェアサイエンス主専攻　201411409　溝口和輝

指導教員　山田武志

提出日2017年8月4日

論文名

DCASE 2017 CHALLENGE SETUP: TASKS, DATASETS AND BASELINE SYSTEM

0. 概要

DCASE 2017 Challenge は、音響シーンの分類、珍しい音響イベントの検出、実生活におけるオーディオのサウンドイベント検出、大規模な車音の監視イベント検出の 4 つのタスクで構成されています。このホワイトペーパーでは、タスク定義、データセット、実験セットアップ、および開発データセットのベースラインシステム結果について説明します。すべてのタスクのベースラインシステムは、多層パーセプトロンとログメルエネルギーを使用する同じ実装に依存しますが、出力レイヤーと意思決定プロセスの構造、タスク固有のメトリックを使用したシステム出力の評価法が異なります。

1. 導入

音は、私たちの日常的な環境や物理的な出来事について多くの情報を持っています。人間は、繁華街、静かな公園、静かなオフィス環境や通過する車、鳥や人の足音などのシーン内の個々の音源を認識する、周囲の環境の一般的な特性を知覚することに非常に慣れています。この情報を自動的に抽出するための計算方法の開発はオーディオコンテンツに基づいたマルチメディア検索[1]、コンテキスト認識モバイルデバイス[2]、ロボットや自動車などのインテリジェントなアプリケーション音響情報を用いて活動を認識するシステム[3][4]などがあります。しかし、環境によって複数の音が存在し、同時に起こるような現実の音風景で、場面や個々の音源を確実に認識するためには、依然として多くの研究が必要です。

DCASE 2017 Challenge では、これまでの成功を踏まえて、一般的に公開されているデータセットを使用してさまざまなアプローチを比較することで、計算シーンとイベント分析メソッドの開発をサポートしています。この分野への継続的な努力は、異なる発展のためのマイルストーンを設定し、さらなる参照のために現在の業績を固定します。課題は、音響シーンの分類、珍しい音響イベントの検出、実生活におけるオーディオのサウンドイベント検出、大規模な車音の監視イベント検出の 4 つのタスクで構成されています。

音響シーンの分類は環境音の分類における重要なトピックです。これは場所や状況の一般的な特性としての音響場面が、一般的な音響特性に基づいて他と区別できるという仮定が存在し、その仮定によって録音された環境を認識するものだとして定義されます。音響シーン分類は DCASE 2013 [5]と DCASE 2016 [6]でタスクとして提示されており、さまざまな方法でアプローチされています。そのために使用される特徴および分類器の一覧としては定番のメル周波数ケプストラム係数[2,8]や、音響イベントのヒストグラム[9]や時間-周波数表現から学習

された勾配のヒストグラムなどのより特化した特徴を含むもの[10]や隠れマルコフモデル(HMM)、ガウスミクスチャーモデル(GMM)、サポートベクトルマシン(SVM)などの音響モデルがあり[7]に提示されています。さらに、深層学習を用いた手法の出現は目覚しいものであり、DCASE2016 の提出されたシステムの多くは様々なタイプの DNN に基づいています[6]。
サウンドイベントの検出は、オーディオ内の個々の音認識として定義され、複数のサウンドクラスの場合もあります。また個別のサウンドイベントインスタンスの開始と終了の推定も含まれます。これは類似の音を単一のクラスとして表すことができると想定しているためであり、このクラスは認識可能にするために他のサウンドクラスとは十分に異なります。音響イベント検出のために最もよく使用される特徴はケプストラル係数または対数エネルギーであるメススケール表現であり、ここには HMM [12]、非負行列分解（NMF）[15,16]、ランダムフォレスト[11]、および DNN が含まれます。
現実のオーディオでのサウンドイベントの検出は、同じサウンドイベントクラスに属するサウンドの固有の音響変化、関わりのあるサウンドイベントと重なる他のサウンドなど、自動メソッドに多くの難点を与えます。いくつかの状況では目的音のみが鳴ることは非常にまれであり、誤検出を避けるために検出システムに付加的な負担を課してきます。 DCASE 2017 では珍しい音響イベントの検出と実生活におけるオーディオのサウンドイベント検出という 2 つの別々のタスクによって、稀なサウンドイベントと高度に重複するサウンドに対応します。
録音された音は分単位のものがインターネット上で共有されます。これらの録音されたものは主にビデオであり、私たちが今までに見た中で最大のアーカイブスを構成しています。これらの音響コンテンツのほとんどはタグなしです。したがって、録音中のサウンドの自動認識はサウンドイベントの検出によって実現できます。しかし文献のほとんどと DCASE のこれまでの 2 回の結果では学習とテストのデータに厳密なラベル（正確なタイムスタンプを含む）が付いているため、音声のみの録音と管理された手法に焦点を当てています。このような注釈を収集高が Web ビデオやサウンドクラスの数に比例することはほとんどありません。したがって我々は、弱いラベル（正確なタイムスタンプを含まない）で学習し、評価される準監督アプローチの必要性があると主張します。現在の文献は、非監督の[18,19,20]や半監督のアプローチ[21,22]、弱いラベルを用いた[23]を用いて潜在的可能性を示しています。このタスクの成功は、ビデオコンテンツ分析のための他の様式を補完するかもしれません。
本稿では、DCASE 2017 のタスクについて詳しく説明します。各タスクについて、タスク定義、データセットに関する情報、タスク設定とベースラインシステム、および開発データセットのベースライン結果を提供します。すべてのタスクのベースラインシステムは同じ実装に依存し、同じ機能と技術を使用します。入力データを処理してターゲット出力にマッピングする方法が異なります。これはアプリケーション固有であり、タスクに従って選択されているためです。

2. セットアップ

DCASE2017 は潜在的な参加者に 4 つのタスクを提供し、公に利用可能なデータセットと各タスクのベースラインシステムを提供しました。チャレンジの出力は、要件に応じてフォーマットされたシステム出力で構成されています。さらに参加者はコミュニティがすべての投稿を比較して理解できるように、システムの記述を含む詳細な技術レポートな提出する必要がありました。課題の締め切りは表 1 に示されており、データセットとベースラインシステムの一般的な構成は以下のセクションで詳細に説明されています。

表 1:チャレンジの〆切

内容	*期日*
開発用データセット解禁	2017/03/21
評価用データセット解禁	2017/06/30
課題の提出	2017/07/31
結果の提出	2017/09/15
ワークショップ	2017/11/16,17

2.1 データセット

チャレンジが開始されたときに、システム開発中に使用される事前に定義された学習とテストセット（クロスバリデーションフォールドフォーマットの一部のタスク）からなる開発データセットが各タスクに提供されました。また開発したシステムの評価のために、評価データセットと呼ばれる別個のデータセットを用意しました。開発データセットはオーディオデータと、タスク固有の形式に関連する注釈と、開発データセットのシステムパフォーマンスを報告するための実験的な設定で構成されています。主催者の提言は、サブミッションを直接比較できるように、提供された実験設定を使用することでした。データセットへのアクセスは、Web サイトを介して提供されました。

すべてのタスクに適用される一般的な規則として、参加者はシステム開発のために外部データを使用することを許可されておらず、異なるタスクのデータセットは外部データとみなされていました。しかし、提供された学習および開発データの操作は、外部データを使用せずに拡張するために許可されました(例えば、データ確率分布関数からのサンプリングや技法を使用してピッチシフトやタイムストレッチを行うなど)。

評価データセットは、チャレンジの提出締め切りの直前に、注釈のないオーディオのみとして提供されました。参加者はこのデータでシステムを実行し、評価のために主催者にシステム出力を提出する必要がありました。参加者は評価データを主観的に判断することも、注釈を付けることもできませんでした。提出されたシステムを訓練するために評価データセットを使用することも禁止されていました。評価データの参照表記は主催者のみが利用できたため、各タスクのメトリックに基づいて結果の評価を実施しました。

2.2 ベースラインシステム

このシステムは、各タスクに対して基本的なアプローチから構成されています。その目的は、参加者がシステムを開発する際に比較ポイントを用意するためです。開発セットのベースラインシステムのパフォーマンスは、各タスクに提供されます。デフォルトのパラメータで実行すると、システムは必要なデータセットをダウンロードし、タスク固有の結果を出力します[24]。

実装は多層パーセプトロンアーキテクチャ（MLP）に基づいており、ログメルバンドエネルギーを音響特徴量として使用します。音響特徴量は周波数範囲 0～22050 Hz をカバーする 40 次元のメルバンドを使用して、40ms のフレームで 50%のオーバーラップで計算されています。特徴ベクトルは、5 フレームの文脈を用いて構築され、200 の特徴ベクトル長になります。MLP は 50 個の中間層を持つ 2 つの層からなり、20%の欠落が存在します。ネットワークは、グラディエントベースの最適化のために Adam アルゴリズムを使用して訓練されます[25]。学習率 0.001 を用いて最大 200 回の学習を行い、100 エポックと 10 エポックの耐久のあとに開始したモニタリングで早期停止基準を使用します。ネットワークの出力レイヤーはタスク固有であり対応するセクションで説明します。ネットワークは前述の機能を使用して訓練され、学習目標は実装されたタスクに従って提示されます。ベースラインシステムには各タスクの特定のメトリックを使用したシステム出力の評価も含まれます。

ベースラインシステムは機械学習用の Keras を使用して Python を使用して実装されました。これにはデータセットの処理、機能とモデルの保存とアクセス、結果の評価に必要なすべて

の機能があり、関連するさまざまなステップの簡単な適応と変更が可能です。参加者には、与えられたベースラインシステムの上にシステムを構築することが許可され、奨励されました。

3. タスク 1：音響シーン分類

音響シーン分類の目的は図 1 に示すように、「公園」「家」「オフィス」など記録された環境を特徴付ける事前に定義されたクラスの中から 1 つを選びテストデータを分類することである。このタスクに提供されるデータセットは、開発セットとしての TUT Acoustic Scenes 2016 [26]と新たに記録された評価セットの 2 つで構成される TUT Acoustic Scenes 2017 です。主な違いは、このタスクでは長さ 3〜5 分の元の録音データが 10 秒の長さのセグメントに分割されており、セグメントは独立した個々のファイルで提供されました。より短いオーディオセグメントは、意思決定プロセスのためにシステムに与える情報が少なくなり、前のエディションよりタスクが難しくなります。この長さは人間と機械の認識の両方にとって困難であるとみなされていて、[2]の研究に基づくものです。データ記録および注釈手順の詳細な説明は[26]に記載されています。

バス、カフェ/レストラン、車、市内中心部、森林路、食料品店、家、湖畔の浜辺、図書館、地下鉄駅、オフィス、住宅街、電車、トラム、都市公園などでした。4 つのフォールドを含むクロスバリデーションの設定が提供され、同じ録音データから得られたすべてのセグメントが学習アルゴリズムの片側に含まれ、トレーニングもしくはテストのいずれかに含まれるように、開発セット内の利用可能なオーディオ素材を分割しました。各クラスについて、開発セットには 10 秒のセグメントが 312 個含まれています。

このタスクではベースラインシステムをマルチクラスのシングルラベル分類設定に合わせて調整されており、ネットワーク出力層は 15 クラスを表すソフトマックスタイプのニューロンで構成されています。分類の決定はニューロンの出力に基づいており、ニューロンの出力は一度に 1 つしかアクティブに出来ません。フレームベースの決定は、過半数の投票を使用して結合され分類されたセグメントごとに単一のラベルを取得しました。システムの性能は、正確なシステム出力の数と出力の総数の比として定義される精度を使用して測定されました[27]。システムは提供された 4 つのクロスバリデーションを使用して訓練及びテストが行われ、開発セット上で 73.8%の平均分類精度を得ました。クラスごとの精度を表 2 に示します。評価セットのシステム性能はカメラ対応バージョンに追加されます。チャレンジのために提出されたシステムの評価は分類精度を用いて行われます。

表 2:ベースラインシステムにおけるクラスごとの精度

音響シーン	*精度[%]*
ビーチ	77.6
バス	83.7
カフェ/レストラン	55.1
カー	86.2
市内中心部	88.5
森林路	83.3
食料品店	63.1
家	74.5
図書館	60.6
地下鉄駅	88.5
オフィス	97.4
公園	64.4
住宅街	62.8
電車	38.1
路面電車	82.7
総計	73.8

4. TASK2：珍しい音響イベントの検出

タスク 2 は図 2 に示すような珍しい音イベントの検出に焦点を当てたものです。 このタスクで使用されるオーディオ素材は人工的に作成された混合音で構成されており、様々なイベントとバックグラウンドの割合で多くの例を作成できます。ここで「まれ」とは半分の録音で最大で 1 回発生する目的の音響イベントを指します。 3 つの目的の音響イベントのクラスついて、これらのイベントの一時的な発生を検出するための別個のシステムを開発します。
提供されたデータセットは、希少なサウンドイベントと背景音の混合音を生成するためのソースファイルと容易に生成された混合音のセット、そして混合音が作成されたいわゆるレシピによって構成されています。さらに追加の混合レシピとオーディオ混合音をさらに生成するソフトウェアパッケージが提供されました。
背景音の録音は TUT Acoustic Scenes 2016 開発データセット[26]に由来していますが、目的のクラスイベントと携帯電話からの干渉を含むセグメントは除きます。まれな音イベントは赤ちゃん叫び（106 トレーニング、42 テストインスタンス、平均持続時間 2.25 秒）、グラスの割れる音（96 トレーニング、43 テストインスタンス、平均持続時間 1.16 秒）、銃撃音（134 トレーニング、53 テスト、平均持続時間 1.32 秒）です。録音データは freesound.org から python wrapper2 で API を介してダウンロード出来ます。このデータセットの「ソース」部分では、これらの録音データは元の形式で提示され、別に分けたイベントの一時的な発生時の注釈が同梱してありました。
我々は freesound.org から取得した完全な長さの録音データから目的の音響イベントを分離しました。これは、実際のイベント、無音領域、および背景音のノイズから構成されます。まず高エネルギーと低エネルギーのフレームを分類するために学習済みの SVM を使用して、半教師付きセグメンテーション[28]を行いました。次いですべての録音データの分析フレームの確率値が上位 10%および下位 10%の加重平均として計算された閾値を用いてアクティブなセグメントを用意しました。そこで得られた各セグメントを注釈をつける人が聴き、関係無いイベントを含むセグメントを破棄しました（赤ちゃんの咳、レーザー銃などの非現実的な響きの銃撃音）。このスクリーニングの過程では、追加でイベントのタイミングの改善がイベントの前後の休止を排除する目的で 100ms 毎に実行され、境界に急激なジャンプは導入されませんでした。
混合音の生成順序は以下のようなパラメータを有します。訓練セットとテストセットの両方の各ターゲットクラスには 500 の混合音がありました。イベントの存在率は 50%です（ターゲットイベントが存在する 250 個の混合音とバックグラウンドのみの 250 個の「混合音」の構成となっていました）。イベント対背景音比（EBR）は、-6、0 および 6dB の 3 種類でした。EBR はイベントの持続時間にわたって計算された平均 RMSE 値と、イベントが混じったバックグラウンドセグメントとの比としてそれぞれ定義された。バックグラウンドインスタンス、イベントインスタンス、混合音中のイベントタイミング、その存在フラグおよび EBR 値は、すべてランダムかつ一様に選択されました。正確な混合音を生成するために必要なデータ（背景音およびサウンドイベントのファイル名、もし存在するならイベントのタイミングおよび振幅のスケーリング係数）は、レシピでコード化されています。レシピは無作為に生成されましたが、乱数発生器の種が固定されているため、再現性があります。
混合音は、レシピにしたがって背景音と対応する目的のイベントの信号を加算することによって生成され、より高いサンプリングレートの場合には合計する前に 44100Hz にダウンサンプリングされます。得られた信号はグローバルな経験的係数として 0.2 でスケーリングされクリッピングを回避しながらダイナミクスを維持しました。その後、量子化ノイズの追加を避けるために、ファイルを 24 ビット形式で保存しました。
このデータセットにはソフトウェアパッケージが付属しています。データセットでデフォルトのパラメータを指定すると、このデータセットとまったく同じ混合レシピと混合音ファイルが生成されます。また、より大規模で困難な訓練データセットを得るためにパラメータを調整することも可能です。パラメータとして混合音の数、EBR 値およびイベントの存在確率は調節可能です。
基礎となるソースデータの観点からトレーニングセットとテストセットに分けるために必要

な情報が提供されました。DCASE 2016 タスク 1 の設定の最初のフォールドに従って 844 個のトレーニングと 277 個のテストファイルを生成し、ロケーション ID の記録によって背景音の分割は行われました。サウンドイベントは freesound.org のユーザー名によって区分されています。目標イベントの例の比率は 0.71：0.29 に設定され、分離されたイベント数が同様の比率であるように分割行われました。その結果の固有のイベントカウントは次のようになります。

・ 赤ちゃん：　　106training 42test;
・グラスの割れる音：　96training 43test;
・銃撃音：　　134training 53test。

ベースラインシステムは共通の実装に従っており、次のような内容になっています。各ターゲットクラスに対して目的のクラスの活動を示し、シグモイド活性化を伴う 1 つの出力ニューロンを有するバイナリ分類器が存在します。ベースラインシステムの挙動は、開発データセット（トレーニングとテストセットからなる）を使用して、イベントベースのエラー率とイベントベースの F-スコアを指標として評価されます。両方のメトリックは[27]で定義されているように 500 ms の襟を使用し、サウンドイベントの開始のみを考慮して計算されます。ベースラインシステムの結果は表 3 に提示されています。評価データセットのシステムパフォーマンスは、カメラ対応バージョンに追加されます。このタスクの主な評価スコアは、イベントベースのエラー率であり、チャレンジのために提出されたシステムのランキングは 3 つのクラスにわたる平均イベントベースのエラー率を使用して行われます。

表 3:タスク 2 のベースラインシステムの性能

イベントクラス	*エラー率*	*F-スコア[%]*
子供の泣き声	0.79	68.1
グラスの割れる音	0.21	89.0
銃撃音	0.72	55.1
平均値	0.57	70.7

5. TASK3：実生活におけるオーディオのサウンドイベント検出

タスク 3 では音源が孤立して聞こえることはほとんどないような、日々の生活と同様にマルチソース状態であるサウンドイベント検出システムのパフォーマンスを評価しました。いくつかのあらかじめ定義されたサウンドイベントクラスが選択され、システムはこれらのサウンドの存在を検出し、図 3 に示すようにテスト音のセグメントにラベルとタイムスタンプをラベル付けすることを意図しています。このタスクでは、学習中もテスト音のデータも、重複するサウンドイベントの数を各時間で制御することはできません。

このタスクに使用されるデータセットは、TUT 音響シーン 2017 のサブセットであり、TUT サウンドイベント 2017 と呼ばれます。さまざまなレベルの交通音や他の活動音を伴う路上（市街地と居住地）の録音データから構成されています。音声の長さは 3〜5 分です。この路上の音響シーンは人間の活動や危険状況に関連する音のイベントを検出するために関わりのある環境を表すものとして選択されました。

[26]に記載されているアノテーションの手順に従って、各録音データの個々の音響イベントには、同じ人が音に対して自由に選びラベルを付けました。音源を特徴付けるために名詞が使用され、音の生成メカニズムを特徴付ける動詞が可能ならば名詞と動詞のペアとして使用されました。ラベルを付ける人は、すべての可聴音響イベントにイベントラベルを自由に付け、適切と思われる音の開始時間と終了時間を決定するよう指示されました。

目標の音響イベントクラスは、人間の存在および交通に関連する共通の音を表すように選択されました。タスクのために選択されたサウンドクラスは、ブレーキ音、車、子供、大型車、会話音、歩行音です。生のラベルマッピングが行われ、音源によって記述されたクラス、例えば「車の通過音」、「自動車の走行中のエンジン音」、「自動車のアイドリング」などの音を「車」に統合し、バスやトラックによるものは「大声車」、「子どもの叫び声」、「子どもの会話音」などが「子供」に分類されています。ラベル付けの過程に個々の主観が高いため、これらのマッピングされたクラスを使用して参照注釈の検証が行われました。 3 人の人（ラベ

ル付けした人を含まない）がこれらのクラスの 1 つに属していると言われた各オーディオセグメントを聴き、そのセグメント内に提示された音が存在するかの確認を取りました。少なくとも 1 人の人は認めたイベントインスタンスが保持され、元のイベントインスタンスの約10%が削除されました。
開発および評価データセットへのデータの分割は、各サウンドイベントクラスで使用可能なサンプルの量に基づいて行われました。録音時の異なるクラスに属するイベントインスタンスが不均等に分散するため、個々のクラスの分割はある程度しか制御できませんが、大部分のイベントは開発セットに含まれます。このデータセットで報告される結果を均一にするために、クロスバリデーションによる評価が提供されています。セットアップは、学習とテストサブセットを含む 4 つのフォールドで構成され、各レコーディングがテストデータとして 1 回だけ使用されるように作られています。クロスバリデーションフォールドを作成する際には、トレーニングサブセットでは使用できないクラスがテストサブセットには含まれていないという条件が課せられました。開発セット内の各イベントクラスのインスタンス数を表 4 に示します。評価セット統計はカメラ対応バージョンに追加されます。
ベースラインシステムは複数のクラスのマルチラベル分類に合わせて調整され、ネットワーク出力レイヤーは同時にアクティブになることができる信号ユニットを含んでいました。この複数の出力はオーバーラップするサウンドクラスのアクティビティを示すことができます。結果は 1 秒のセグメント長を使用して、セグメントベースのエラー率とメトリックとしてセグメントベースの F-スコアを使用して評価されます。4 つのクロスバリデーションフォールドは 1 つの実験として扱われ、メトリックは個々のフォールディングやクラスの挙動を平均化するのではなく、すべてのフォールドでエラーカウント（置換、削除、置換）の累積によってメトリックが計算されます[27]。この計算法はクラスのバランスとエラータイプの影響を受けず、セグメント毎のサウンドインスタンスに等しい重みを与えます[29]。表 5 に示すように、提供されたクロスバリデーションを使用してシステムを学習させ、テストした結果、総合エラー率は 0.69、全体 F スコアは 56.7%でした。完全性のために、個々のクラスの結果が全体の結果に沿って示されています。評価データセットのシステムパフォーマンスは、カメラ対応バージョンに追加されます。このタスクの主な評価スコアは全体的なセグメントベースのエラー率であり、チャレンジのために提出されたシステムのランク付けも評価データセットで計算された同じメトリックを使用して行われます。

表 4:タスク 3 の開発用セットにおけるクラスごとの総数

イベントラベル	*ラベル数*
ブレーキ音	52
車	304
子供	44
大型車両	61
歩行音	109
会話音	109
総計	659

表 5:タスク 3 の結果(セグメントベース)

イベントクラス	*エラー率*	*F-スコア[%]*
ブレーキ音	0.98	4.1
車	0.57	74.1
子供	1.35	0.0
大型車両	0.90	50.8
歩行音	1.25	18.5
会話音	0.84	55.6
全体値	0.69	56.7

6. TASK4：大規模な車音の監視イベント検出

タスク 4 は、弱いラベルを付けされたオーディオ録音を使用してサウンドイベントの大規模な検出のためのシステムを評価するものです。音声は交通と警告のトピックに関する YouTube の動画抜粋からのものです。トピックは業界の関連性とこの文脈でのオーディオの不足のために選ばれました。この結果は大規模なサウンドイベント検出の新しい根拠を定義し、自動運転車やスマートシティおよび関連分野における音響の利点を示すのに役立ちます。このタスクは 10 秒間のクリップ内のサウンドイベントを検出することで構成され、2 つのサブタスクに分割されました。

・ サブタスク A：タイムスタンプなし（同じスキャンタグ付き、図 4）
•サブタスク B：タイムスタンプあり（タスク 3 と同様、図 3）

タスクはオーディオセット [30]のサブセットを使用しました。オーディオセットは 632 のサウンドイベントクラスのオントロジーと、YouTube のビデオから集められた 200 万人がラベルを付けた 10 秒間のサウンドクリップのコレクションで構成されています。オントロジーはイベントカテゴリの階層グラフとして指定されていて、人間や動物のさまざまな音、楽器や音楽ジャンル、一般的な日常的な環境音をカバーします。データセットを収集するために Google は人間の注釈者と協力して、YouTube の 10 秒クリップで聞いた音を聞き、分析し、確認しました。すべてのクラスのサンプルをすばやく蓄積するために、Google は利用可能な YouTube メタデータとコンテンツベースの検索を頼ってターゲットサウンドを含む可能性の高い候補の動画セグメントを探しました。 オーディオセットには 10 秒間のクリップ内の各サウンドクラスに正確な時間境界が設定されていないため、ラベルは弱いものとみなされています。また、1 つのクリップは複数のサウンドイベントクラスに対応してもよいものとみなします。タスク 4 は警告音と車両音の 2 つのカテゴリに分けられた 17 のサウンドイベントのサブセットから成立しています。

・ 警告音：トラックの音、電車の音、車の警告音、ビープ音、救急車のサイレン、警察のサイレン、消防車のサイレン、民間警備車のサイレン、叫び声
・ 車両音：自転車、スケートボード、車、車の行き交う音、バス、トラック、オートバイ、電車

どちらのサブタスクでも、データは 2 つの主な区切りで分けられました。開発用と評価用です。開発データはそれ自体が学習用とテスト用に分けられていました。トレーニングには 51,172 個のクリップがあり、クラスごとにアンバランスでサウンドイベントごとに少なくとも 30 個のクリップがありました。テストにも 488 個のクリップがあり、クラスごとに少なくとも 30 個のクリップがありました。10 秒のクリップは複数のサウンドイベントクラスに対応している可能性があります。評価セットには 1,103 個のクリップがあり、サウンドイベントごとに少なくとも 60 個のクリップが含まれていました。このセットにはオーディオ内の特定のサウンドイベントが存在することを示す程度のラベルがありましたが、タイムスタンプの注釈はありませんでした。テストと評価のために、サブタスク B のパフォーマンス用の強

いラベル（タイムスタンプ注釈）が提供されています。タスクのルールで他のデータセットなどの外部データを使用することは禁止されています。同じように残りのビデオサウンドトラック、ビデオフレーム、メタデータ（テキスト、ビュー、お気に入り）など、10 秒クリップが抽出されたビデオの他の要素を使用することもできませんでした。さらに参加者は、オーディオセットや外部データを間接的に使用する埋め込み機能を使用することも禁止されていまし。システムの学習には弱いラベルと強ラベル（タイムスタンプ）の注釈だけが使用することを許されていました。

2 つのサブタスクの評価は異なっていました。タイムスタンプなしのサウンドイベント検出（音響タグ付き）のために F-score（精度とリコール）を使用し、提示されたシステムのランキングは F スコアに基づくものです。タイムスタンプを使用したサウンドイベント検出ではセグメントベースのエラー率[27]と F スコアが使用され、ここでは提出されたシステムのランキングが 1 秒セグメントベースのエラー率に基づいていました。

ベースラインシステムはコードの元を他のタスクと共有し、同時にアクティブにすることができるシグモイドユニットを含むネットワーク出力レイヤーに基づいて検出を行います。ベースラインには結果の評価も含まれ、表 6 に示されています。ベースラインシステムはトレーニングセットを使用して学習され、テストセットを使用してテストされました。評価セットの詳細とシステム性能は、カメラ対応バージョンに追加されます。

表 6:Task4 のサブタスク A,B のベースライン実行時の結果

クラス	*サブタスク A[%]*			*サブタスク B*	
	F-スコア	*精度*	*リコール*	*エラー率*	*F-スコア[%]*
電車の警笛音	0.0	0.0	0.0	1.00	–
トラックの警笛音	11.9	72.4	6.5	1.00	–
車のクラクション	6.8	10.4	5.1	1.00	–
ビープ音	5.9	53.8	3.1	1.00	–
救急車	37.3	57.1	27.7	1.00	–
パトカー	49.5	43.1	58.3	0.96	19.1
消防車	26.3	25.7	27.0	1.02	1.2
民間警備車のサイレン	55.2	42.0	80.6	0.71	55.7
叫び声	37.7	60.1	27.5	1.00	1.0
自転車	36.4	49.3	28.9	1.00	–
スケートボード	20.7	70.8	12.1	1.00	–
車	9.5	5.0	100.0	4.34	20.2
車の行き交う音	5.2	29.1	2.8	1.00	–
バス	17.0	18.7	15.6	1.00	–
トラック	25.3	21.8	30.3	1.00	–
バイク	36.5	74.1	24.2	1.00	–
電車	9.2	16.4	6.3	1.00	–
全体値	19.8	16.2	25.6	1.00	11.4

7. 結論

DCASE 2017 チャレンジでは、環境音の分類における現在の研究に関連する 4 つの課題が提案されました。これまでのチャレンジと比較して、現在の版ではごくまれにしか表れない音響イベントの検出と、弱いラベルを用いた音響イベント検出システムの問題という特定の 2

種類の状況に取り組んでいます。確立されたタスクの中で、現実の音声における音響シーン分類および音響イベント検出は重要であると見なされていますが、進行中の研究を含めるにふさわしい問題を解決するにはまだ早いです。
公的なデータセットと結果の報告を通して、このチャレンジはオープンな研究と出版を促進し結果を多くのオーディエンスに広めるかもしれません。提供されたベースラインシステムは、各タスクの設定と比較の基準と共にさらなる開発の出発点として提供しました。

8. 出典

[1] *M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search." in Interspeech, 2009, pp. 1151–1154.*

[2] *J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audiobased context recognition," IEEE Trans. on Audio, Speech,*

[3] *D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustic (WASPAA), October 2015.*

[4] *S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J. Appell, and ¨ F. Wallhoff, "Acoustic monitoring and localization for social care," Journal of Computing Science and Engineering, vol. 6, no. 1, pp. 40–50, March 2012.*

[5] *D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," IEEE Trans. on Multimedia, vol. 17, no. 10, pp. 1733–1746, October 2015.*

[6] *T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, and M. Lagrange, Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016). Tampere University of Technology. Department of Signal Processing, 2016.*

[7] *D. Barchiesi, D. Giannoulis, D. Stowell, and M. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 16–34, May 2015.*

[8] *J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-offrames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," The Journal of the Acoustical Society of America, vol. 122, no. 2, pp. 881–891, 2007.*

[9] *T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in 18th European Signal Processing Conference, Aug 2010, pp. 1272– 1276.*

[10] *A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 23, no. 1, pp. 142–153, Jan. 2015.*

[11] *B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording," in DCASE2016 Workshop on Detection and Classification of Acoustic Scenes and Events, 2016.*

[12] *A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in 18th European Signal Processing Conference (EUSIPCO 2010), 2010, pp. 1267–1271.*

[13] *J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in 22st ACM International Conference on Multimedia (ACM-MM'14), Nov. 2014.*

[14] *E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291–1303, June 2017.*

[15] *J. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach to audio event detection," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Oct 2013, pp. 1–4.*

[16] *A. Mesaros, O. Dikmen, T. Heittola, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 151–155.*

[17] *E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM*

neural network," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2015.

[18] B. Byun, I. Kim, S. M. Siniscalchi, and C.-H. Lee, "Consumer-level multimedia event detection through unsupervised audio signal modeling." in INTERSPEECH, 2012, pp. 2081–2084.

[19] B. Elizalde, G. Friedland, H. Lei, and A. Divakaran, "There is no data like less data: Percepts for video concept detection on consumer-produced media," in Proceedings of the 2012 ACM international workshop on Audio and multimedia methods for large-scale video analysis. ACM, 2012, pp. 27–32.

[20] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. Jackson, and M. D. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1230–1241, 2017.

[21] W. Han, E. Coutinho, H. Ruan, H. Li, B. Schuller, X. Yu, and X. Zhu, "Semi-supervised active learning for sound classification in hybrid learning environments," PloS one, vol. 11, no. 9, p. e0162075, 2016.

[22] A. Shah, R. Badlani, A. Kumar, B. Elizalde, and B. Raj, "An approach for self-training audio event detectors using web data," arXiv preprint arXiv:1609.06026, 2016.

[23] A. Kumar and B. Raj, "Weakly supervised scalable audio content analysis," in 2016 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2016.

[24] T. Heittola, A. Diment, and A. Mesaros, "DCASE2017 baseline system," https://github.com/TUT-ARG/ DCASE2017-baseline-system, accessed June 2017.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2014.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in 24th European Signal Processing Conference 2016 (EUSIPCO 2016), 2016.

[27] ——, "Metrics for polyphonic sound event detection," Applied Sciences, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

[28] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," PloS one, vol. 10, no. 12, 2015.

[29] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," SIGKDD Explor. Newsl., vol. 12, no. 1, pp. 49–57, nov 2010. [Online]. Available: http://doi.acm.org/10.1145/1882471.1882479

[30] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in Proc. IEEE ICASSP 2017, New Orleans, LA, 2017.