# 抽象

本稿では、自動サウンドイベント検出（SED）タスクのための長期短期記憶（LSTM）リカレントニューラルネットワーク（RNN）と組み合わせた空間的および調和的特徴の使用を提案する。実生活のサウンドレコーディングは、通常、重複するサウンドイベントがあり、モノラルチャンネルのオーディオだけでは認識しにくいものです。人間の聴取者は、ピッチキューを使用してオーバーラップするサウンドイベントの混合を認識し、これらのイベントを空間的に特定するために耳にあるステレオ（マルチチャネル）オーディオ信号を利用しています。伝統的に自動音声イベント検出システムは、マルチチャンネルオーディオを使用するようにそれらを拡張することを提案する人間の聴取者によって動機付けされたモノラルチャンネルオーディオのみを使用してきた。提案された自動音声イベント検出システムは、TUTサウンドイベント検出2016データベース[1]の開発サブセット上の最先端のモノチャンネル方法と比較される。提案手法は、誤り率を6％削減しながらFスコアを3.75％改善する。

*インデックス規約*-サウンドイベント検出、マルチチャンネル、到着時間dif-フェレンス、ピッチ、リカレントニューラルネットワーク、長期短期記憶

# 1.はじめに

サウンドイベントは、人間の聴取者が音響環境でも一貫してラベル付けして区別できるオーディオのセグメントです。このような自動音声イベント検出（SED）のアプリケーションは数多くあります。リスニング機能を備えた組み込みシステムは、環境をより認識しやすくなります[2]。産業界および環境監視システム、およびスマートホームは、関心のあるイベントを自動的に検出することができます[3] [4]。マルチメディアの自動注釈は、コンテンツベースの照会方法のためのより良い検索を可能にすることができる[5] [6]。

自動音声イベント検出の課題は、連続するオーディオ信号のサウンドイベントを認識することです。これまでに構築されたサウンドイベント検出システムは、モノラルとポリフォニック(多声的なもの)に大別できます。単音システムは、音声信号の中で最も支配的なサウンドイベントを認識するように訓練されている[7]。ポリフォニック(多声的な)システムは、最も支配的なサウンドイベントを超え、セグメント内のすべての重なり合ったサウンドイベントを認識します[7] [8] [10] [11] [12]。本稿では、このような実在のシナリオを再現するような多音な音景に取り組むことを提案する。

いくつかの自動音声イベント検出システムは、メル係数周波数ケプストラム係数（MFCC）と隠れマルコフモデル（HMM）を用いて、ビタビアルゴリズムを連続的に通過する分類器としてポリフォニック検出に取り組んできた[7]。[13]では、非負値行列因子分解(NMF)を前処理ステップとして使用し、各ストリームの中で最も顕著な事象が検出された。しかし、重複イベントの数を見積もるという厳しい制約がありました。これは、[14]で結合されたNMFを使用することによって克服された。Dennisら[8]は、一般化されたハフ変換（GHT）と局所スペクトル特徴を組み合わせることによって、伝統的なフレームベースの特徴とはまったく異なる経路をとった。

より最近では、最先端の自動音声イベント検出システムは、マルチラベル分類のために訓練されたDNN [11]およびRNN-LSTM [12]ネットワークにおいて、ログメルバンドエネルギーの特徴を使用してきた。[12]に示されているように、DNNに対するRNN-LSTMの良好な性能に動機づけられ、我々は引き続きマルチラベルRNN-LSTMネットワークを使用する。

現在の技術の多声自動音声イベント検出システムは、サウンドイベント検出のために単一チャネルのオーディオを使用してきた。多チャンネルのデータを持っていれば、ポリフォニックイベントが潜在的に改善される可能性があります。人間が2つの耳（2つのチャンネル）を使用して周りのサウンドイベントを認識してローカライズするのと同じように、多チャンネルオーディオからのサウンドイベントを学習するようにマシンをトレーニングする可能性もあります。近年、Xiaoら[16]は、遠隔地の自動音声認識（ASR）のためにマルチチャネル音声から空間特徴をうまく使用し、単なる音声チャネルを使用することに比べてかなりの改善を示した。これは、自動音声イベント検出タスクのために空間的特徴を使用することをさらに促す。本稿では、アクセント特性とともに空間特性を提案し、60分程度の小さなデータセットでもモノチャネル特性よりも優れていることを示す。

論文の残りの部分は次のように構成されています。第2節では、使用された特徴と提案されたアプローチを説明する。第3説では、RNNと長期短期記憶（LSTM）ブロックについて簡単に紹介します。第4節では、実生活の記録データベースを実験的に設定し、結果を報告する。最後に、5章で結論を提示する。

# 2.サウンドイベントの提案手法

サウンドイベント検出タスクは、サウンドイベントの位置を時間的に特定し、それらを既知のラベルセットの1つに割り当てることを目指す。実生活の健全な出来事には特別なパターンはありません。森林、都市、家庭などのさまざまな状況では、さまざまな種類のサウンドイベントがあります。コンテキストに基づいて異なる希少性を持つことができ、単独で発生することも、他のサウンドイベントと完全に重複することもあります。孤立した音はかなりの精度で認識されているが[17]、重ね合わされた音のイベントでのラベルの混合を検出することは困難な作業であり、かなりの量の改善が可能である。図2は、音声イベント注釈のスニペット(≒フレーム)を示しています。ここでは、会話音、車、および犬の鳴き声の3つのイベントが発生します。時間フレーム*t*では、2つのイベント-会話音と車がオーバーラップしています。理想的な自動音声イベント検出システムは、そのような重なり合った事象を処理することができるべきである。

人間の聴覚システムは、音声イベントを分離、ローカライズ、分類するために、耳に届くステレオ（マルチチャンネル）オーディオ情報をうまく利用しています。サウンドイベント検出システムがステレオ入力を取得し、サウンドイベントをローカライズおよび分類するために適切な空間的特徴が実装される同様のセットアップが想定され、実施される。

図1に示す提案されたサウンドイベント検出システムは、現実のマルチチャネルオーディオレコーディングで動作し、孤立したオーバーラップするサウンドイベントを検出し、分類することを目的としています。

ステレオオーディオから、3つの特徴 - ログメルバンドエネルギー、ピッチ周波数、およびその周期性、およびサブバンドにおける到着時間差（TDOA）が抽出される。すべての特徴は、特徴間で一貫性を持たせるために、20 msのホップ長で抽出されます。

## 2.1.Logmel-bandEnergy

モノラルチャンネルのサウンドイベントの検出には、ログのメルバンドエネルギーが広く使われており[11] [12] [18]、優れた特徴であることが証明されています。提案されたシステムでは、ログメルバンドエネルギーを引き続き使用し、両方のステレオチャンネルに対してそれを抽出する。これは、人間の聴覚システムが、音源の空間的位置特定のための両耳間レベル差（IID）を利用するという考えから動機づけられている[15]。ニューラルネットワークは線形演算を実行することができ、その差異を含む。したがって、ステレオのログメルバンドエネルギーデータを訓練するとき、両耳間レベル差IIDと同様の情報を得ることを学ぶ。

オーディオの各チャンネルは[ハミング窓関数](https://ja.wikipedia.org/w/index.php?title=%E7%AA%93%E9%96%A2%E6%95%B0&action=edit&section=11)を使用して50％のオーバーラップを有する40msのフレームに分割される。その後、ログメルバンドエネルギーが各フレーム（表1のmel）について抽出される。私たちは全スペクトルにわたって40メルバンド帯を使用しています。

## 2.2.ハーモニー特性

ピッチは音の重要な知覚的特徴です。ヒューマン・リスナーは、ピッチ・キューを使用して異なる音を識別するように進化しており、オーバーラップするサウンド・イベントで混合音のそれぞれを音響的に分離するためにピッチを効率的に使用できます[19]。Uzkentら[20]は、MFCCに加えてピッチ範囲を用いた非会話の環境音検出の精度の改善を示している。ここでは、絶対的なピッチとその周期性を特徴として使用することを提案する（表1のピッチ）。

スレッショルドなパラボリック補間STFT [22]のピッチトラッキング[21]のlibrosa(ライブラリ名)実装を使用してピッチと周期性を推定しました。

私たちはマルチラベル分類を扱っているので、できるだけ多くの有意な基本周波数を特定し、それを使ってサウンドイベントを特定するのは直感的です。周期性の特徴は、抽出されたピッチ値の信頼度を示し、分類器がピッチに基づいてより良い決定をするのを助ける。

トレーニングデータ（第4.1説）のオーバーラップするサウンドイベントは、一度に3つ以上のイベントが重複していなかったため、フレームごとに主要な3つのピッチ値を使用するように制限されています。したがって、各チャネルについて、上位3つのピッチ値、およびそれぞれの周期性値が、100~4000Hzの周波数範囲（表1のピッチ3）の各フレームで抽出される。

## 2.3.到着時の時間差（TDOA）機能

オーバーラップするサウンドイベントは、分類システムに問題を与えます。主に、重なったフレームの特徴ベクトルが異なるサウンドイベントの組み合わせであるためです。しかし、人間の聴取者は、空間的に音源を分離し、定位することによって、重なり合ったサウンド事象の各々を首尾よく識別することができた。これは、両耳間時間遅延（ITD）のためにのみ可能であった[15]

各サウンドイベントにはそれぞれ独自の周波数帯があり、低周波にはいくつか、高周波数にはいくつか、周波数帯にはすべてが発生します。周波数スペクトルを異なる帯域に分割し、これらの帯域のそれぞれにおける音源の空間的位置を特定することができれば、これは特徴の別の次元であり、分類器はそれぞれの可能な源の数を推定することを学ぶことができるフレーム、およびその方向を空間内で決定する。我々は、スペクトルフレームを5つのmel-bandsに分割し、これらの帯域のそれぞれにおける到着時間差（TDOA）を計算することによってこれを実施する。

例えば、重なり合わない独立したサウンドイベントが全周波数範囲に広がっている場合、TDOAを5つのメルバンドで計算しています。バンドごとに同じTDOA値を設定する必要があります。我々は2つの重複した音S1とS2を持っている場合には、S1は最初の二つの帯域に拡散され、S2は後半の二つの帯域に拡散されています。特徴ベクトルは、音のそれぞれについて異なるTDOA値を有し、分類器はそれらを分離して別個のサウンドイベントとして識別することを学ぶことができる。

TDOAは、位相ベース重み付け（GCC-PHAT）[23]を用いた一般化相互相関を用いて推定することができる。ここで、各メルバンドの相関を別々に抽出する。

Rb(Δ12,t)=Hoge …(1)

Nは、周波数帯域の数であり、X(k,t)は、時間フレームtにおけるk番目の周波数帯域のFFT係数であり、Hb(k)はB帯域合計のb番目のメル帯域の振幅反応であり、Δ12は、チャンネル間のサンプルのディレイ値です。TDOAは、各メルバンドおよび時間フレームの相関ピーク強度の位置として抽出される。

τ(b,t)=argmax{Rb(Δ12,t)}

最大値と最小TDOA値は2τmaxから-2τmaxの間に切り納められていて、τmaxはマイク間を移動する音波間の最大サンプル遅延である。トレーニングセットのサウンドイベントは、50msから数秒まで変化することがわかった。このような可変長サウンドイベントに対応するために、TDOAは3つの異なるウィンドウ長（120,240,480 ms、一定ホップ長20 ms）で計算されました。これらの3つのウィンドウのTDOA値は、1つのTDOAフィーチャセットを形成するために、各メルバンドに関連付けられていました。したがって、5つのメル帯域で抽出されたTDOA値、および3つのウィンドウ長について連結された場合、フレーム当たり15のTDOA値が得られる（表1のtdoa3）。

小さなウィンドウ内のTDOA値は、一般的に非常にノイズが多く、信頼性がありません。これを克服するために、フレームの各サブバンドについての上記3つの異なるウィンドウ長からのTDOA値の中央値をTDOA特徴の第2のセット（表1のtdoa）として使用した。ウィンドウ長にわたってフィルタリングをポストし、各メルバンドのTDOA値も、アウトライアを除去するために長さ3のカーネルを使用して時間的に中央値フィルタリングされた。

# 3.マルチレーベル支援ネットワーク・ベースのイベント・イベント検出

深いニューラルネットワークは、音声認識[24]、画像認識[25]、機械翻訳[26]など、複雑なパターン認識タスクで非常にうまく機能することが示されている。深いニューラルネットワークは、通常、いくつかの後続の行列乗算および非線形活性化関数を介して、入力から出力空間へのマップを計算します。モデルのパラメータ、すなわちその重みとバイアスは、勾配降下のような最適化の形を使って反復的に調整されます。

ネットワークが有向非循環グラフである場合、すなわち、情報が前方伝播のみである場合、それはフィードフォワードニューラルネットワーク（FNN）として知られている。フィードバック接続がある場合、モデルはリカレントニューラルネットワーク（RNN）と呼ばれます。RNNは、以前のタイムステップの情報を隠れたレイヤに組み込むことができ、オーディオタスクの時間的コンテキストなどのシーケンシャルデータに基づいてタスクのコンテキスト情報を提供します。長時間短期記憶（LSTM）[27]などの複雑なRNNアーキテクチャは、消失勾配問題を軽減するために近年提唱されている[28]。LSTMは現在、RNNの最も広く使用されている形式であり、この作業でも使用されています。

児童音声イベント検出では、RNNを使用して、タイムスタンプtにおいて、各クラスが所定のフレームでアクティブである確率を予測することができる。ネットワークへの入力は、一連の特徴ベクトルx（t）である。ネットワークは各隠れ層の隠れアクティベーションを計算し、出力層では各クラスy（t）の予測ベクトルを計算する。出力層では、複数のクラスを同時にアクティブとして予測できるように、シグモイド活性化関数が使用されています。出力層で予測をしきい値処理することにより、バイナリ・アクティビティ・マトリクスを得ることが可能である。

## 3.1。ニューラルネットワーク構成

各記録について、特徴ベクトルのシーケンスを取得し、それをゼロ平均および単位分散に正規化し、テスト特徴ベクトルを正規化するためにスケーリングパラメータを保存する。シーケンスはさらに、5フレームのホップ長を有する長さ25のより小さいシーケンスに分割される。これらのフレームのそれぞれは、どのクラスが特徴ベクトル内に存在するかを示すターゲットバイナリベクトルを有する。

我々は、それぞれが32個のLSTMユニットを有する2つの隠れ層を有するマルチラベルRNN-LSTMを使用する。入力レイヤーのユニット数は、使用されているフィーチャーの長さによって異なります。出力層は、各クラスに対して1つのニューロンを有する。ネットワークは、ロス関数としてのバイナリクロスエントロピーとAdamオプティマイザ[30]を使用して、時間経過によるバックプロパゲーション（BPTT）[29]によって訓練される。早期停止を使用してオーバーフィッティングを減らすと、検証セットのセグメントベースエラーレート（ER）（セクション4.2を参照）が100エポックタイムで減少しない場合、トレーニングは中断されます。

テスト時に、トレーニングデータで推定されたスケーリングパラメータを使用して、特徴ベクトルをスケーリングし、それらを25フレームの重なり合わないシーケンスで提示し、固定閾値が0.5である出力を閾値化する、すなわち、ネットワークの出力レイヤのポストリライアンスは0.5より大きく、そうでない場合は非アクティブです。

# 4.評価と結果

## 4.1。データセット

我々は、提案された児童音声イベント検出システムを、TUTサウンドイベント検出2016データベースの開発サブセット[1]で評価する。このデータベースには、バイノーラルSoundman OKM II Klassik / studio A3エレクトレットインイヤー型マイクとRoland Edirol R09ウェーブレコーダーを使用して44.1 kHzサンプリングレートと24ビット分解能でステレオ録音を行いました。それは2つの文脈、すなわち家庭と居住区を含む。ホームコンテクストは11のサウンドイベントクラスを含む10のレコーディングを持ち、住宅エリアコンテクストは7つのクラスを有する12のレコーディングを有する。これらの録音の長さは3〜5分です。

提供されている開発サブセットでは、各コンテキストデータは、トレーニングデータとテストデータの4つのフォールドに既に分割されています。テストデータは、各記録がテストとして正確に1回使用されるように収集され、その中のクラスは常にトレーニングデータのクラスのサブセットです。また、各折りたたみにおける訓練データ記録の20％を、妥当性確認データとしてランダムに選択した。同じ検証データがすべての評価で使用されました。

## 4.2。メトリック

我々は、[31]で定義されたサウンドイベント検出のために確立されたメトリックを使用する[1]と同様の方法で、システムの評価を実行する。エラー率（ER）およびFスコアは、1秒間の長いセグメントで計算されます。すべての折り目の結果が結合され、単一の評価が生成されます。これは、[32]で議論されているように、折り目間のデータ不均衡に起因するバイアスを避けるために行われます。

## 4.3。結果

データセット[1]のベースラインシステムでは、40 msフレームと20 msホップ長のモノオーディオで抽出された20個の静的（0番目の係数を除く）、20個のデルタと20個の加速MFCC係数が使用されます。次に、16個のガウス分布からなるガウス混合モデル（GMM）をクラスの正と負の値ごとに訓練する。このベースラインシステムは、コンテキスト平均ERが0.89、Fスコアが26.6％である。理想的なシステムは、1のERと100％のFスコアを有するべきである。

表2では、提案された空間的特徴および調和的特徴の異なる組み合わせに対するセグメントベースのERおよびFスコアを比較する。これらのすべての評価では、フィーチャセットに基づいて入力レイヤーのサイズのみが変更され、RNN-LSTMネットワーク内の残りの設定は変更されません。

モノチャンネルの特徴を備えたマルチチャンネルオーディオの提案された空間的および調和的な特徴の性能を比較するために、我々はステレオチャンネルを平均化することによってモノラルチャンネルオーディオを作成した。これらのモノラルチャンネルのオーディオでは、セクション2.1と同様の方法でログのメルバンドの特徴が抽出されました。RNN-LSTMネットワークはこれらの機能を訓練し、システムを評価しました。文脈を越えて、Fスコアは、ERを変化させないGMMベースラインシステムより良好であると見られた。ここでは、このモノラル・ログ・ログ・メル・バンド機能とRNN-LSTMネットワーク構成の結果を比較のベースラインとして使用します。

表2にMEL1以外の全ての組み合わせを示すように、ハイブリッドな組み合わせのセットが試された。TDOAは、F-スコアにコンテキスト間のベースラインよりも良好に行います。それらのすべてがERに関してベースラインを下回っていた。最後に、RNN-LSTMネットワークとのさまざまな組み合わせで提案された空間的および調和的特徴の全スペクトルを評価した。例外のカップルと-メル2;tdoa3及びMEL2;tdoa3、ピッチ2、特徴の組み合わせの全てが基準としてわずか同様の平均ERと、等しいまたは平均Fスコアにおけるベースラインよりも良好に形成パー。MEL2の結果。tdoa;ピッチ2機能の組み合わせは、コンテキスト間で際立っていると、与えられたデータセットとネットワーク構成のための最高の特徴ベクトルであることを見られています。

# 5。結論

本論文では、RNN-LSTMネットワークと併せて、マルチラベルサウンドイベント検出に空間高調波特徴を用いることを提案した。評価は、60分の限定されたデータセットサイズで行われました。このデータセットには、家庭と居住地域の2つの状況に関する4つの交差検証データが含まれています。提案されたマルチチャネル機能は、モノチャネル機能を使用するベースラインシステムよりも大幅に優れていることがわかりました。

将来の研究は、新規のデータ増強技術の発見に集中する。スペースフィーチャを拡張することは未踏のスペースであり、検討する価値のある課題になります。モデルに関しては、それらを双方向RNNに拡張し、畳み込みニューラルネットワークと結合するように、RNNの異なる構成についてさらなる研究を行うことができる。

HOGE

これらの結果につながる研究は、ERC助成契約637422 EV-ERYSOUNDの下、欧州研究評議会からの資金提供を受けています。フィンランドのCSC-ITセンター（フィンランド）に計算資源を提供したいと考えています。

表2：家庭および居住地域のコンテキストにおける異なる機能の組み合わせで達成されたセグメントベースのエラーレート（ER）およびFスコア。表1に示す機能は、提案されたRNN-LSTMネットワークとは異なる組み合わせで使用されます。下付き文字「1」及び「2」の特徴の組み合わせ列のは、機能がオンに抽出されたどのように多くのチャンネルを表します。例えば、特徴の組み合わせMEL2;TDOA;ピッチ2は、最終的な特徴ベクトルは、メルバンドエネルギーをログ両方のステレオチャンネルに抽出最も支配的なピッチ及び周期性の値を有し、かつ到着の時間差（TDOA）の間で計算することを意味しますステレオチャンネル。強調表示された各コンテキストのERとFスコアのペアは、達成された最高のERスコアです

https://www.gstatic.com/images/branding/googlelogo/1x/googlelogo_color_48x16dp.png

**英語の原文テキスト:**

Motivated by the good performance of RNN-LSTM over DNN as shown in [12], we con- tinue to use the multi-label RNN-LSTM network.

http://www.google.com/images/zippy_plus_sm.gif翻訳を改善する