M-015

学習済み音声イベント分類器を用いた音響情報の系列解析に基づく居住空間の状況推定手法

A Method for Estimating Residential Situations Based on Sequential Analysis of Acoustic Information Using a Pretrained Audio Event Classifier

湯本 稜矢 †　和崎 克己 ††
Ryoya Yumoto　Katsumi Wasaki

1　はじめに

AI や IoT 技術の発展により，カメラやセンサーを使った生活者の状況を推定する手法が存在する．スマートホームの分野においても活躍しており，日常生活の効率化や安全性の向上が期待されている．一方で，そういったモニタリングやセンシング技術を用いた行動推定手法には，プライバシーの懸念や設置・装着のわずらわしさなどの侵襲性が課題となっている．本研究では住居空間で発生する生活音・環境音に着目した，音声分類技術を活用した非侵襲的な生活状況を推定する手法である．住居空間で発生する音には，その空間を利用する者の日常生活における様々な状況や，状況の組み合わせに関する情報が含まれており，音の収音・分析を通じて利用者の行動や状況を明らかにすることができる．

2　環境音分類の学習済モデル（YAMNet）

YAMNet は音声イベント分類器で，音声波形を入力とし，AudioSet オントロジー [1] で定義された 521 クラスの音声イベントそれぞれに対して独立した予測を行う．モデルには MobileNet v1 アーキテクチャを使用しており，AudioSet コーパス使用してトレーニングされている．環境音分類や音響イベント検出などのタスクに使用でき，計算量が小さく非常に軽量であるため，今回のような，家庭内の環境音分類のリアルタイム処理が必要な場合に適している．このモデルは,[-1.0,+1.0] に正規化されたモノラル 16kHz サンプルで表される波形を入力とする．入力された波形は，内部的に長さ 0.9 秒とホップ 0.48 秒のスライド窓でフレーム化し，これらのフレームのバッチに対してモデルのコアを実行する．出力の一つである scores は各フレームに対する 521 種類のクラスそれぞれの予測スコアを格納した，(N,521) の形状を持つ float32 のテンソルである [2].

YAMNet で分類される音声イベントは具体的なものが多く，状況・雰囲気のような抽象的な概念を表すものは少ない．状況・雰囲気を推定するには単一の音声波形では不十分であり，後述のように音声スライスを単位とした確率遷移モデルによる推定が必要と考えた．

3　生活状況推定手法

本研究では，日常生活で発生する可能性の低い音声イベントは対象から除外した．また，YAMNet で分類できる音声イベントを音の発生源に基づいて「人」「物」「環境」のカテゴリに分け，危険な音声イベントを選定した．以下のようなフローで生活状況を推定する．

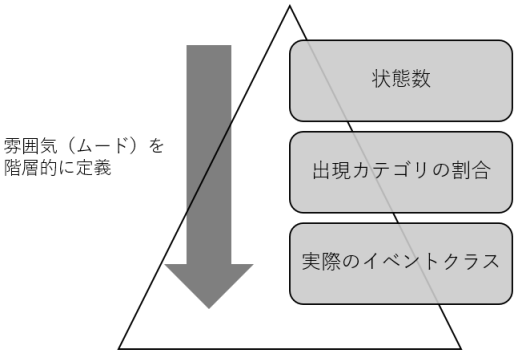

図 1　雰囲気の階層的な推定モデル（状態数，出現カテゴリの割合，実際のイベントクラスの順番で抽象度を下げる方向で階層化して定義）

1. **音声データの前処理・分割**　短期的な環境の変化や状況を捉えるために，前処理されたデータを例えば 10 秒間を単位としたセグメントでスライスする．生活状況を把握するための短期的な変化をとらえるのに十分な時間を確保する．
2. **各セグメントの分類・カテゴリや危険ラベル情報の付加**　YAMNet により各セグメントのイベントクラスとその尤度を算出し，上位の 2 つの音声ラベルを抽出して順不同で組み合わせ，各セグメントの状態とする．これにより，音声データを人間が理解できる具体的な状況として捉えられる．あわせて，事前に振り分けたカテゴリや危険音ラベルもここで付与する．
3. **状態遷移確率の計算・グラフ化**　得られた分類結果を基に状態間の遷移確率を計算し確率状態遷移モデルとして得る．この解析により，状態がどのように遷移するのかが明らかとなり，状態の変化パターンやシーケンスを確認できる．状態遷移の視覚化は，生活者の行動や環境の変化を理解するために重要であり，特に生活空間における異常や特定の状況の発生を予測するための手がかりとなる．音声分類結果の時間的関係を定量的に示すことで，生活者の状態や環境変化をより深く理解できる．
4. **状態数やカテゴリなどに基づいた雰囲気の推定**　得られた結果を解析し，状態の総数や音声分類結果に基づいて雰囲気を推定する．雰囲気の推定モデルは，図 1 に示すように状態数，出現カテゴリの割合，実際のイベントクラスの順番で抽象度を下げる方向で階層化して定義する．

† 信州大学大学院総合理工学研究科, Graduate School of Science and Technology, Shinshu University
†† 信州大学工学部電子情報システム工学科, Department of Electrical and Computer Engineering, Faculty of Engineering, Shinshu University

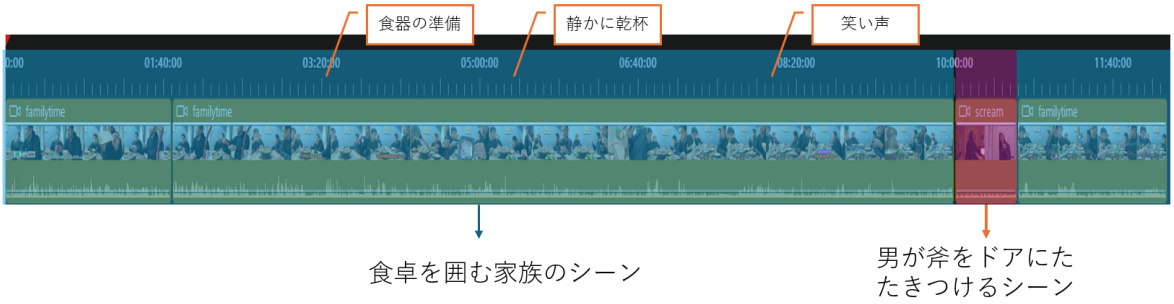

図 2　試行実験に用いた YouTube 動画のタイムライン（12 分．家庭内での食事シーンがほとんどで，後半の数十秒間に異常なシーンを加工・挿入したもの）

4　システム概要、試行実験

4.1　システム概要

YAMNet の分類精度と提案手法の妥当性を検証するため，音声ファイルの読み込みから前処理，分割，音声分類，グラフ化，雰囲気の推定までを一貫して実施する分析プログラムを実装した．YAMNet には，tensorflowHub[3] から取得した学習済モデルを利用する．作成する確率状態遷移モデルは，状態と遷移確率に加え，クラス名を音声カテゴリ別に色分けし，危険な音を検知したノードを視覚的に識別できるように工夫した．試行実験には，YouTube から取得・編集した動画音声を用い，図 2 のように手動で簡易的なラベルを付与し，その結果と分類結果を比較して精度を評価した．

雰囲気推定は条件分岐により実施し，階層的な定義に基づいて上位レベルから順に雰囲気を表す語を選び，最終的に自然なフレーズへと構成した．この際，状態数・カテゴリ比率・支配的状態に応じた言葉の選択条件を設けている．

4.2　試行実験

今回の試行実験によって得られた確率状態遷移モデルを図 3 に示す．以下に推定結果について概要を分類スコアの観点から述べる．

「Speech_Music」ノードに着目すると次の点が確認できた．約 90% の確率で自身の状態に遷移し，他状態への遷移はわずかであった．遷移先には，「Speech_Laughter」や「Music_Dishes, pots, and pans」が含まれ，危険な音としては「Door_Slam」と「Arrow_Thump, thud」が検出された．

実際の動画との比較すると，「Speech_Music」は動画内での話し声と，BGM が該当しており，高い確率での自身への遷移は動画の内容と一致していた．「Speech_Laughter」や「Music_Dishes, pots, and pans」について，これらはそれぞれ，07:40 の笑い声と 03:25 の食器の準備に該当していた．05:40 の乾杯音に該当する状態は見られなかった．「Speech_Domestic animals, pets」について，動画内にペットは出現しておらず，該当箇所はなかった．分類スコアを見ると「Speech: 0.85, Domestic animals, pets: 0.01」という結果であり，比較すると Speech のスコアが高かった．危険な音として検出した「Door_Slam」，「Arrow_Thump, thud」については，斧のシーンが合致することが確かめられた．

状態数，出現するカテゴリの比率，支配的な状態から，

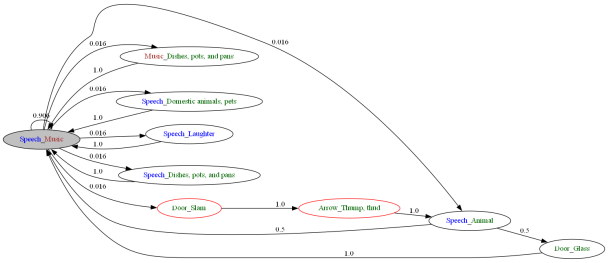

図 3　確率状態遷移モデルの例

推定される雰囲気は “落ち着いていてやや活気のある Speech_Music 空間” となった．この雰囲気は，動画の内容と一致しており，妥当な結果であると考えられる．

5　まとめと今後の課題

試行実験の結果は良好であったが，他方，分類スコアに極端な偏りがあると誤分類されるケースもあった．ソフトマックス関数の影響でスコアを奪いあった影響だと考察する．スコアに閾値を設けて結果を取得する対応が必要がある．

試行時に用いた動画内には BGM と話し声のほか，瓶を置く音や箸と食器の接触音，YouTube としての効果音も含まれていた．これらは極めて短時間であり，約 12 分の動画に対して手動ラベリングの対象とするのは困難であったため省略した．結果としてこれらの音は分類にも現れず，推定された雰囲気も妥当であったことから，短い音の影響は限定的であると考えられる．

一方で異常検出の目的を優先するためには，短い音が分類されにくいということは推定精度が劣化する原因になりうるため，取得する音声データの単位時間については検討が必要である．

雰囲気の推定に関しては，少し漠然としているように感じられた．雰囲気をより具体的に推定するためには，今後，推定可能なラベルに関して，さらに細分化したり，雰囲気定義の階層を増やしたりする必要がある．また，雰囲気の定義方法については他分野からのアプローチも視野に入れての検討が必要である．

参考文献

[1] AudioSet. (2023). AudioSet Ontology. https://research.google.com/audioset/ontology/index.html
[2] kaggle. (2023). yamnet. https://www.kaggle.com/models/google/yamnet
[3] TensorFlow. (2023). Sound classification with YAMNet. https://www.tensorflow.org/hub/tutorials/yamnet