視聴覚マルチモーダルAIの開発と移動空間評価への適用

曽　翰洋[1]・葉　健人[2]・土井　健司[3]・中島　悠太[4]

[1]学生会員　大阪大学大学院　工学研究科地球総合工学専攻（〒565-0871 大阪府吹田市山田丘2-1）
E-mail:sou.kanyou@civil.eng.osaka-u.ac.jp

[2]正会員　大阪大学大学院助教　工学研究科地球総合工学専攻（同上）
E-mail:yoh.kento@civil.eng.osaka-u.ac.jp

[2]正会員　大阪大学大学院教授　工学研究科地球総合工学専攻（同上）
E-mail: doi@civil.eng.osaka-u.ac.jp

[4]非会員　大阪大学大学院教授　産業科学研究所（〒567-0047 大阪府茨木市美穂ヶ丘8-1）
E-mail: n-yuta@im.sanken.osaka-u.ac.jp

自動車優先から人中心への転換が求められる都市空間デザインにおいては，移動空間を単なる通行のための空間から，滞留や活動を支える場へと再定義し，都市環境により形成される利用者印象を計画・設計に反映する理論的枠組が不可欠である．本研究では，72カ国にわたる歩行映像から抽出した約15,000セグメントに主観評価に基づくアノテーションを施し，視聴・聴覚環境などの空間属性を結び付け，統合的に処理するマルチモーダル AI モデルを構築した．この結果，視覚・聴覚単独に比べ，視聴覚を統合したモデルはより高い推定精度を達成した．モデルの出力結果から，各環境音カテゴリに対する空間性能評価の分布を分析し，両者の関係性を明らかにするとともに，感度分析を通じて視聴覚情報に基づく評価モデルとしての妥当性と適用可能性を示した．

Key Words: *pedestrian friendly, urban perception, multimodal AI, audiovisual, AI evaluation*

1. はじめに

都市の成熟化に伴い，自動車優先から人中心への空間再編が求められている．こうした都市空間の転換においては，移動空間を単なる通行のための空間から，滞留や多様な活動を支える場へと再定義する必要がある．高齢化社会[1)]や気候変動への対応[2)]，持続可能な都市への転換といった社会的要請により，効率性よりも人々の健康や交流[3)]を支え，生活の質（QOL）や Well-being を高める空間の質が重視されるようになっている[4),5)]．

こうした背景から，歩行者優先の都市環境整備が国際的にも加速している．バルセロナのスーパーブロック政策[6)]では車両通行を制限し歩行者・自転車優先空間を創出し，パリの15分都市構想[7),8)]では日常機能を徒歩圏内に配置している．メルボルンの20分生活圏政策[9)]でも地域内での歩行促進により社会的つながりの強化が進められており，これらはいずれも車両中心から徒歩圏での生活への転換を通じて，持続可能で質の高い都市空間の実現を目指している．日本でも，国土交通省による「居心地が良く歩きたくなるまちなか」づくりが推進されており，歩きやすさ(Walkable)，まちへの開放性(Eye level)，多様性(Diversity)，開かれた空間(Open)を重視した人中心の空間形成が進められている[10)]．これらの政策に共通するのは，道路や街路を通行機能に特化した空間から，人々の交流や活動を支える公共空間へと転換する視点である．

このような転換を実現するには，人々にとって居心地が良く活動を促す空間特性を定量的に評価する手法が必要である．従来の都市空間評価は交通容量，歩行者密度，道路幅員といった機能的指標が中心であった[11),12),13)]．しかし，人々の空間利用行動は，物理的な空間構成だけでなく，その空間で得られる経験に強く影響される．実際，都市空間の様相は利用者の主観性に影響を与え，行動や健康などの

観点から生活の質と因果関係を有することが明らかになっている[14),15),16)].

都市における空間経験は，視覚的景観，音環境などの多様な感覚情報を通じて空間への印象を形成し，滞留や交流等の行動を誘発する一連のプロセスとして捉えられる[17),18),19)]．したがって，空間質を評価するには，利用者の都市環境に対する知覚や印象を定量的に評価し，その知見を計画・設計に反映する理論的枠組みが求められる．

本研究では，このような背景を踏まえ，都市空間における利用者の印象を視覚および聴覚情報から推定するマルチモーダル AI を開発し，空間性能評価への適用可能性を検証することを目的とする．具体的には，世界各国の歩行空間動画から視覚情報と聴覚情報を抽出し，居心地の良さや賑わいといった空間性能指標との関係を学習することで，環境音を含む多様な感覚的要素が空間的経験に与える影響を定量的に評価する手法を提案する．

2. 先行研究レビュー

(1) 都市空間の視覚的な印象評価

都市空間における主観的印象評価では，従来，アンケートや現地調査といった手法が用いられてきたが，これらはデータ収集に多大なコストを要する[20)]とともに，都市全体にわたって高い空間解像度で主観的評価を実施することは困難であった．この課題に対し，近年では，地理ビッグデータや Web 調査をベースとした AI 評価手法が提案されつつある．具体的には，ストリートビューから取得可能な街路景観画像と，クラウドソーシングを活用した主観的評価データの収集手法が注目されている．これらの手法では，ウェブ調査を通じて大量の街路景観画像に対する主観的評価データを効率的に紐付け，収集されたデータに基づいて印象評価を推定する機械学習モデルや深層学習モデルが開発されている．

Naik ら[21)]は，大規模オンライン調査により，Google Street View 画像に対して，安全性評価を収集し，サポートベクター回帰により知覚された安全性を予測する StreetScore を開発した．Quercia ら[22)]は，ロンドンの街路景観における美しさ，静けさ，幸福度から都市の美的資本を定量化した．さらに，Salesses ら[23)]は，欧米の 4 都市において大規模なペア比較評価を収集し，都市の知覚的不平等を測定する手法を提案した．

これらに対し，Dubey ら[24)]は Place Pulse （56 都市，110,988 枚の画像，1,170,000 件のペア比較）を用いて，安全性，活気，退屈さ，豊かさ，憂鬱さ，美しさの 6 つの知覚次元について深層学習モデルによる予測手法を開発した．Naik ら[25)]は，時系列の Street View 画像を用いて都市の物理的変化を定量化し，教育水準や人口密度が近隣環境の改善を予測する要因であることを明らかにした．

これらの基盤的研究に続いて，モデルの解釈性や適用範囲の拡張を目指した研究が展開されている．Rossetti ら[26)]は，Place Pulse データセットに離散選択モデルを適用し，景観要素と主観的知覚の関係をより解釈可能な形で明らかにした．大規模都市圏への適用として，Zhang ら[27)]や Wei ら[28)]は深層学習モデルを用いて，大規模都市圏における都市景観の人間知覚を高精度にマッピングする手法を開発した．また，都市計画への実装を志向した研究として，Wang ら[29)]は深層学習と Space Syntax を統合することで住民の街路知覚を測定する定量的評価手法を提案している．さらに，Kang ら[30)]は，GeoAI による街路景観画像からの安全知覚評価と地域住民調査による安全知覚評価を比較し，両手法の特性と限界を明らかにした．評価次元の多様化という観点では，街路景観画像から 22 種類の主観的知覚を同時に測定する深層学習ベースの分類モデルも開発されている[31)]．また木崎ら[32)]は，経路選択モデルと画像解析技術，生成 AI を組み合わせることで，街路景観が歩行者の経路選択行動に及ぼす影響を定量的に評価し，街路空間整備の施策評価へ適用している．

(2) 都市空間の聴覚的な印象評価

都市空間における人間の環境認知は，視覚情報だけでなく聴覚情報にも大きく影響される．歩行者は移動中に交通騒音，自然音，人々の活動音など，多様な音環境に曝されており，これらの聴覚情報は空間の快適性や安全性の知覚に影響を及ぼすことが知られている[33)]．特に，騒音は歩行者の心理的快適性や健康に直接的な影響を及ぼし，静穏性の高い空間は歩行行動を促進する要因となる[34)]．本節では，都市空間における聴覚的な印象評価に関する研究動向を整理する．

従来の研究では，心理実験やアンケート調査を通じて人々の音環境に対する主観的な評価構造を明らかにする試みがなされてきた．これらの研究により，音環境の知覚には「快適さ」や「活気」といった複数の評価次元が存在すること[35),36)]，個人差や文脈によって音源の認識パターンが異なること[37)]，音環境の時空間的変動が景観特性や都市活動と関連していること[38)]などが明らかにされている．こうした知見の蓄積を踏まえ，騒音低減中心のアプローチ

から人間の音体験を積極的にデザインする方向への転換が提唱されている[39].

一方，AI 技術の進展により，従来人手に依存してきた音環境分析が大規模かつ自動的に行えるようになりつつある．深層学習による環境音分類では，Piczak[40]が畳み込みニューラルネットワークを用いて従来手法を上回る精度を示した．また，UrbanSound8K データセットを用いた実験では，Long Short-Term Memory を用いて，高精度な音源識別が可能となったことが報告されている[41),42]．実用的な応用として，AudioSet データとベイズ最適化ニューラルネットワークを用いた都市街路の音源マッピング[43]や，音源識別と不快度予測を同時に行うモデル[44]などが提案されている．

これらに対し，視覚情報から音環境を推定する新たなアプローチも提案されている．これらの手法では，コンピュータビジョン技術により街路画像から視覚特徴を抽出し，機械学習モデルによって音環境指標を予測する．Zhao ら[45]は，機械学習モデルによって 15 種類の音環境指標を予測する手法を提案した．シンガポールと深センを対象とした分析では，約 50 万枚のストリートビュー画像に対して予測を行い，現地調査との比較によりその有効性を確認している．同様に，Rui ら[46]は福州市において 98,000 枚の画像を用いて，機械学習アルゴリズムによる音環境予測の有効性を示している．これらの手法は，従来の音響測定では困難であった都市全域の高解像度の環境音評価を可能にする．

(3) 視覚情報と聴覚情報の統合

前節までに述べたように，都市空間の印象評価に関しては，主に視覚情報に基づく景観評価や聴覚情報に基づく音環境評価が個別に進められてきた．しかし，都市における空間経験は多感覚反応に基づくものであり[47]，空間に対する印象評価を実態に即して捉えるためには，視覚と聴覚の双方を統合して扱う視点が不可欠である．たとえば VR 空間における主観評価実験では，両モダリティの組み合わせが都市空間の自然性や満足度の知覚に相互的な影響を与えることが報告されている[48].

これを受け，視聴覚情報を統合的に扱う試みが展開されつつある．Verma ら[49]は，機械学習を用いて手動収集された視覚・聴覚データセットから都市知覚を理解する手法を提案した．さらに同グループは，ムンバイの混在用途地域において街路景観画像と音声クリップを時空間的に収集し，深層学習モデルによって高・低レベルの特徴を抽出することで，視覚的・聴覚的知覚の予測モデルを個別に構築し，時空間的な知覚地図を作成した[50].

より包括的なアプローチとして，クロスモーダルな視点から都市感知を行う研究も提案されている．Chen ら[51]は，ロンドン，ニューヨーク，東京の 3 都市において，ジオタグ付き音声記録とストリートレベル画像・リモートセンシング画像を用いて，音と視覚情報の整合性を評価する Cross-Modal Urban Sensing の手法を開発した．この研究では，事前学習された埋め込みモデルを用いてクロスモーダル類似度を評価し，ストリートビューの埋め込み表現が環境音との整合性において優れていることを示した．また，Chen ら[52]は，福州の大学キャンパスにおいて視聴覚環境と注意回復の質の関係を機械学習により解明し，景観設計への応用可能性を示した．

これらの研究は，視聴覚情報の統合が都市空間の印象評価において重要であることを示している．しかし，これらの研究にはいくつか限界が存在する．第一に，多くの研究では，視聴覚情報を個別に処理した後に統合する段階的アプローチを採用しており，両モダリティを統合的に処理する方法論の構築は限定的である．第二に，研究対象が単一都市または限定的な地域に留まっており，グローバルスケールでの視聴覚環境の多様性を捉えていない．第三に，先行研究は主に予測精度の向上に焦点が置かれており，環境音が空間性能に与える影響といった，両モダリティ間の関係性分析は十分に行われていない．

本研究では，72 カ国にわたる歩行映像から抽出した 15,523 セグメントに基づき，視聴覚情報を統合的に処理するマルチモーダル AI モデを構築する．視覚・聴覚単独モデルとの比較を通じて視聴覚統合の有効性を検証するとともに，各環境音カテゴリに対する空間性能評価の分布分析と感度分析を通じて，視聴覚情報に基づく評価モデルの妥当性と適用可能性を明らかにする．

3. 方法論

本研究では，都市空間の性能評価を目的とし，視聴覚情報を統合するマルチモーダル AI を開発する．まず，Web 上から CC ライセンスが付与された歩行空間の動画を収集し，ノイズ除去等の前処理を行う．次に，各動画に対してアノテーションを実施し，都市空間に対する印象および環境音カテゴリのラベルが付与されたデータセットを構築する．その後，環境音および景観画像のラベル付きデータを入力情報として，VGG16 ベースのマルチモーダル AI のアーキテクチャの設計および学習を行う．最後に，学習

済みモデルを用いて都市空間の性能評価を行い，空間性能と環境音の関係を明らかにするとともに，感度分析を通じて視聴覚マルチモダリティに基づく空間性能評価の有効性を検証する．

(1) データセット構築

対象データである歩行空間の動画は，主にYouTubeからCCライセンスの付与された動画を収集した．多様な都市空間に対する印象を反映するため，世界各国から幅広い歩行動画を収集することを目的とした．具体的には，検索クエリとして英語で「都市名_walking video」を用い，各国における人口第三位までの都市名を対象とした．また，日本を対象とした歩行動画に関しては，日本語および英語の両言語で「都道府県名_walking video」，「都道府県名_歩行動画」をキーワードとして検索を実施した．その結果，世界の各地域から合計250本の動画が収集された．具体的には，アジア（日本除く）から48本，オセアニアから1本，北米から22本，中南米から13本，欧州から108本，中東から11本，アフリカから9本，さらに日本から38本が得られている．各動画の長さは10分台から2時間台まで多様であり，総計約69時間の映像データとなる．

次に先行研究の知見を踏まえ，本研究では各動画を6秒のセグメントに切り出した．Aumondら[53]は瞬間的な音の心地よさの評価は過去6秒間の平均音圧レベルに影響を受けることを示唆している．Chuら[54]は，2秒から6秒のオーディオクリップを用いた環境音認識において，6秒のクリップの平均認識率が85%と最も高く，識別精度の向上に寄与することを示している．Vermaら[43]は深層学習モデルを用いた環境音分類において，10秒間の環境音サンプルには，不要なノイズ要素や他クラスと類似した誤分類されやすいサウンドイベントが含まれやすいため，正確な判定が困難になると指摘している．これらを踏まえ，本研究では各動画を1セグメント6秒として，約42,000セグメントに分割した．なお，連続して抽出されたクリップには重複情報が含まれるため，一定の間隔で抽出するフィルタリング処理を施し，最終的に22,500セグメントをアノテーション用データセットとして採用した．

また本研究では，評価作業の一貫性と効率化を図るために，回答用UIやユーザーマニュアルなどの詳細なガイドを含む専用のwebアプリケーションを開発し，各評価者がブラウザ上で直接アノテーション作業を実施できる環境を整備した．本アプリケーションでは，画面上の左側にセグメント動画の再生ウィンドウが配置され，右領域に評価項目が提示される構成となっている（**図-1**）．

都市交通計画の専門知識を有する3名のアノテーターが，それぞれ7,500本のセグメント動画を対象として，都市空間に関する空間性能評価および環境音の分類を行った．空間性能評価では，都市空間の「Lingerability[55),56)]：居心地の良さ」と「Vibrancy：賑わい」について，1（最低評価）から4（最高評価）の4段階で印象を評価した．環境音については，交通音，活動音，自然音の3つのカテゴリに分類した[57),58)]．各セグメントにおいて複数の音要素が存在する場合，最も支配的な音を選択し，音の識別が困難な場合は「無音」として評価するよう指示した．各指標およびカテゴリの定義を**表-1**に示す．なお，シーン転換や動画編集，または撮影上の不具合等により，都市空間の印象や環境音を適切に評価できないと判断された映像は，アノテーションの際に評価対象外のセグメント動画として除外している．

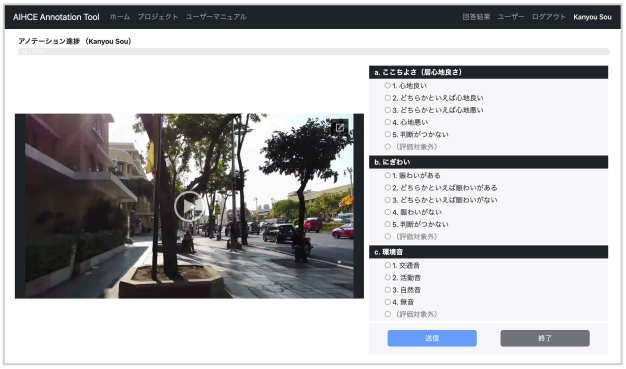

図-1 アノテーション用Webアプリケーション

表-1 アノテーション項目の定義

空間性能評価 （印象評価）	
居心地良さ Lingerability	空間に佇み，とどまりたくなる居心地の良さに加え，快体験の余韻を楽しむためのゆったりとした移動を促す性質
賑わい Vibrancy	その空間に活気があり，楽しいといったようなポジティブな感情を促す性質
環境音分類	
交通音	自動車，二輪車，大型車，バス，鉄道，飛行機，サイレン音,クラクション，アイドリング音など，公共交通や自動車運行に起因する音
活動音	会話，足音，子供の声，工事音，工場音，店内音楽，広告音，機械音など，人の活動に関連する音
自然音	鳥のさえずり，動物の鳴き声，昆虫やカエルの音，風の音，水の音，葉擦れの音など，自然由来の音

アノテーションにおいて評価対象外と判断されたセグメントを除外した結果，最終的に22,500本のうち，15,523 本のセグメントを AI モデルの学習用データとして採用した．各セグメントについて実施した「居心地良さ」と「賑わい」の 2 指標に対する 4 段階の印象評価，および環境音の 4 カテゴリ分類のアノテーション結果の分布を**表-2** に示す．

「居心地良さ」の評価では，1〜4 の各段階における割合はそれぞれ 11.2%，22.8%，30.4%，35.6%であり，特に評価 1 に該当するデータの割合が相対的に少ないことが確認された．「賑わい」の評価では，評価 1〜4 の割合は順に，27.0%，22.7%，27.7%，22.7%とおおむね均等に分布していた．「居心地良さ」については，クラス間の不均衡が顕著であったため，学習データの比率およびモデルの安定性を考慮し，評価 1 と 2 を統合するかたちで，4 クラスから 3 クラスへ再構成した．また，モデル設計の一貫性を保つために，「賑わい」についても同様に評価 1 と 2 を統合し，3 クラス構成とした．

環境音の分類においては，交通音が 30.3%，活動音が 58.9%，自然音が 9.9%，無音が 0.9%であった．無音に分類された 132 本は，音の識別が困難であると判断されたため，分類タスクから除外し，交通音・活動音・自然音の 3 クラスによる分類を行った．

表-2 各ラベルにおけるアノテーション分布

	評価 4	評価 3	評価 2	評価 1
居心地良さ	5532 (35.6%)	4712 (30.4%)	3542 (22.8%)	1737 (11.2%)
賑わい	3520 (22.7%)	4297 (27.7%)	3520 (22.7%)	4186 (27.0%)
	交通音	活動音	自然音	無音
環境音	4706 (30.3%)	9154 (58.9%)	1531 (9.9%)	132 (0.9%)

(2) マルチモーダル AI のアーキテクチャ

本節では，マルチモーダルAIの構築手法について詳述する．**図-3**に示す通り，本研究で提案するマルチモーダルAIは，視覚および聴覚の各モダリティに対して個別の畳み込みニューラルネットワーク（CNN）を適用し，最終的に全結合層において各特徴量を統合する構成を採用している．具体的には，各都市空間の歩行動画から映像および環境音の情報を抽出し，それぞれを2次元画像データに変換する．

映像データは動画の先頭フレームを抽出した上で，画像を224×224ピクセルにリサイズし，各画素値を0〜1の範囲に正規化した．一方，環境音データは，音声信号をメルスペクトログラムへ変換した後，同様に224×224ピクセルへのリサイズおよび正規化処理を施した．メルスペクトログラムは，周波数軸を人間の聴覚特性に基づくメル尺度で表現したものであり，人の可聴範囲における音の知覚特性や心理音響学的な知見[59]を反映している．また，メルスペクトログラムを用いた音響分類が従来のMFCCベースの手法より高い精度を示すことが報告されており[61]，画像分類モデルとの親和性も高い．このため，主観的な都市の音環境評価を分類問題として扱う本研究に適した特徴表現といえる．

モデルの学習には，事前学習済モデルとしてVGG16[61]を採用し，ファインチューニング[62]を適用して最適なパラメータを学習する．ファインチューニングとは，事前学習済みモデルの最終出力層である全結合層を新たな出力層に置き換え，後段の畳み込み層の重みを再学習する手法である．CNNでは，入力側の浅い層が色やエッジなどの汎用的な特徴を抽出する一方，後段の層は学習データに特化した抽象的な特徴を獲得する傾向がある．このため，浅い層の重みを固定し，後段の重みのみを目的タスクに合わせて再学習させることで，スクラッチで学習す

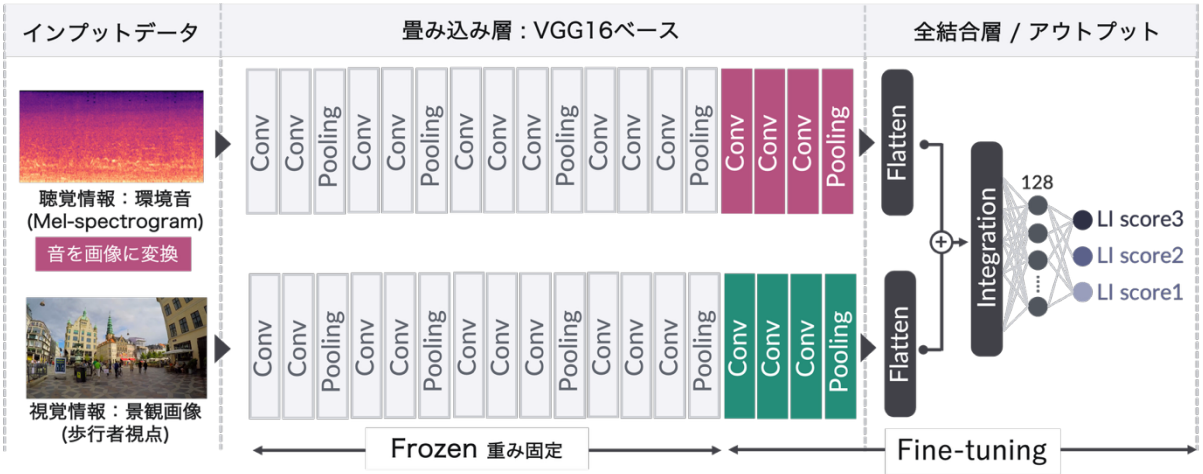

図-3 マルチモーダル AI の構造

るよりも効率的かつ高精度なモデルの構築が可能となる．VGG16は，ImageNetで事前学習されたCNNモデルであり，13層の畳み込み層と3層の全結合層，計16層から構成される．結果として，幅広い画像に対する豊富な特徴表現を獲得しており，ファインチューニングに適した汎用的な事前学習済みモデルとして知られている[63]．

そこで，各モダリティに対してVGG16のBlock5以降の畳み込み層を再学習対象とするファインチューニングを適用し，都市空間の印象評価および環境音分類に適した特徴量抽出を実現する．全結合層では，各CNNから得られた特徴量を1次元の特徴ベクトルへ変換した上で，結合することによって統合的な特徴表現を構成する．最終的には，この統合された特徴量に基づき，印象評価または環境音の分類を行う．印象評価においては，各都市空間データに対して，Lingerability（居心地良さ）およびVibrancy（賑わい）のスコア1〜3に対応する順序付きカテゴリの確率分布を推定する．一方，環境音分類においては，交通音，活動音，自然音の3カテゴリに対する確率分布を推定する．

データセットは，全データに対して各クラスのサンプル数が均等になるように10%をテストデータとして抽出した．残りの90%については，層化K分割交差検証を適用し，5つのグループに分割する．具体的には，交差検証の各反復において，1つのグループ内で80%を訓練データ，残り20%を検証データとして用い，全てのグループが一度は検証データとなるように5回の繰り返しを行った．

学習にあたっては，クロスエントロピーを損失関数として採用し，Adam オプティマイザを用いてパラメータの最適化を行なった．学習率やバッチサイズといったハイパーパラメータの設定は，検証データを用いた探索的実験により調整した．その結果，学習率 5×10^{-4}，バッチサイズ 32，最大エポック数 30 を採用した．訓練中に検証損失が一定期間改善しなかった場合には，性能停滞時に学習率を自動的に減衰するようにし，学習の安定性と効果的な収束を図った．具体的には，検証損失が 5 エポック改善しない場合に学習率を 0.1 倍に減衰させる設定とした．また，ドロップアウト（率 0.6）や L2 正則化（λ=0.01），早期終了（patience=10 エポック）などの正則化手法を導入し，過学習を抑制することでモデルの汎化性能を向上させている．さらに，クラス不均衡に対処するため，各クラスのサンプル数に応じた重み付けを損失関数に適用した．

(3) 都市空間評価への適用

マルチモーダルAIにより出力される順序付きカテゴリの確率分布をもとに，Lingerability index（LI）およびVibrancy index（VI）を算出し，都市空間の性能評価の指標とする．具体的には，AIモデルの確率分布の出力に対して，スコア変換することにより，定量的な指標を導出する．スコア変換においては，各クラスi（i=1,2,3）にスコアS_i（S_i=1,2,3）を割り当て，モデルが出力するクラスごとの確率分布P_iを重みとする加重平均を用いて算出し，指標の値が0から1の範囲に収まるように正規化を行う．

$$X = \sum_{i=1}^{N} P_i S_i \tag{1}$$

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

環境音と空間性能の関係を明らかにするため，学習に用いていない 1000 本の移動空間のセグメント動画に対して，マルチモーダル AI を適用し，LI スコアおよび環境音カテゴリ（交通音・活動音・自然音）を推定する．これにより算出した LI スコアを 0.1 刻みの区間に分割し，各環境音カテゴリに分類されたデータを母数として，各スコア区間に該当するデータの割合（クラス内割合）を算出する．これにより，居心地の良さの評価における環境音の分布特性を定量的に明らかにする．

さらに，環境音が空間性能評価に与える影響を定量的に検証するため，感度分析を実施する．具体的には，学習に用いていないデータから歩行者中心空間(Pedestrian-Centric space: PC)および車両中心空間(Automobile-Centric space: AC)を代表する街路画像をそれぞれ 100 件抽出し，各画像に対して元の環境音を異なる音環境へ置換することで，音環境の変化による評価値の変動を分析する（**図-4**）．

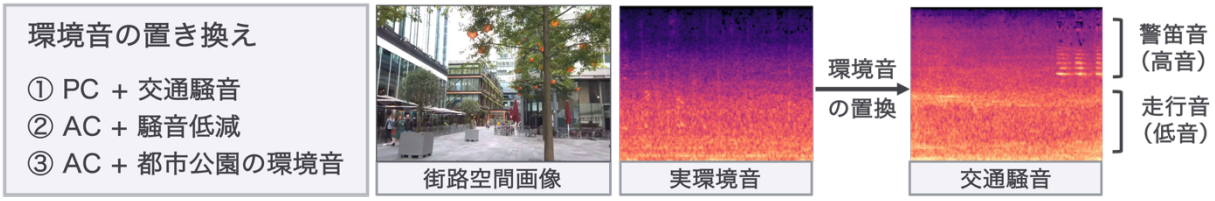

図-4 感度分析のための環境音置換イメージ

置換に用いる環境音として，交通騒音，騒音低減，都市公園の環境音の 3 種類の代表音声を設定する．1)PC の画像と交通騒音，2)AC の画像と騒音低減環境音，3)AC の画像と都市公園の環境音をそれぞれ組み合わせることで，歩行者空間での騒音曝露，交通静穏化，歩車共存の各シナリオを表現し，合計 300 の仮想的な視聴覚シナリオを構成している．

各ケースについて，マルチモーダル AI を適用して LI と VI を算出し，元の環境音との組み合わせ時の評価値と比較することで，環境音の置き換えに対する評価の感度を定量的に分析する．

4. 研究結果と考察

(1) AI モデルの学習結果と精度比較

本研究で構築したマルチモーダルモデルおよび，比較対象として構築した視覚情報のみ，聴覚情報のみを入力とするシングルモーダルモデルの学習結果を**表-3** に示す．表中には，Lingerability（居心地の良さ），Vibrancy（賑わい），環境音の 3 指標に対して，訓練データ，検証データ，テストデータにおける正解率を記載している．

まず Lingerability に関して，マルチモーダルモデルのテストデータにおける正解率は 0.56 であり，視覚（0.54），聴覚（0.46）と比較して最も高い精度を示した．視覚モデルと聴覚モデルの正解率の比較から，空間の「居心地の良さ」の推定には，視覚情報がより有効であることが示唆される．Vibrancy に関しても同様にマルチモーダルモデルが 0.67 と最も高い推定精度を示した．シングルモーダルモデルにおいては，視覚モデル（0.61）と聴覚モデル（0.62）の間には大きな差は見られなかった．このことから，賑わいに対しては，視覚と聴覚情報いずれも同程度に寄与していることが読み取れる．

なお，両指標に関して，シングルモーダルモデルは訓練データでいずれも正解率 1.00 を示している一方で，テストデータでは大きく精度が低下している．これは，シングルモーダルの場合，学習データへの過適合が生じやすく，汎化性能が低いことを示唆している．対照的に，マルチモーダルモデルは訓練データでの精度はやや抑えられているが，テストデータでは相対的に高い精度を維持しており，過学習の抑制と汎化性能の向上が確認できる．

環境音に関しては，マルチモーダルモデルが 0.82 と最高値を示した．聴覚モデル単独でも 0.80 と高い正解率を示している一方，視覚モデルは 0.64 に留まっている．環境音ラベルは，当然ながら聴覚的特徴に基づいて付与されているため，音響情報の有無が推定精度に大きく影響していると解釈できる．

以上より，いずれの指標においてもマルチモーダルモデルは視覚または聴覚の単独モデルを上回る推定精度を達成しており，視聴覚情報を統合的に処理する意義が確認された．特に Lingerability や Vibrancy といった抽象的な印象指標に対しては，視覚情報と聴覚情報を併用して処理することでより安定かつ高精度な推定が可能となることが示された．

続いて，マルチモーダルモデルに対する 3 指標の混同行列を**図-5** に示す．左から順に，Lingerability，Vibrancy，および環境音の分類結果を示している．

表-3 マルチモーダル AI モデルの学習結果とシングルモーダルとの精度比較

モダリティ	Lingerability			Vibrancy			環境音		
	訓練	検証	テスト	訓練	検証	テスト	訓練	検証	テスト
マルチモーダル	0.77	0.52	0.56	0.82	0.65	0.67	0.99	0.82	0.82
シングルモーダル（視覚）	1.00	0.75	0.54	1.00	0.74	0.61	0.96	0.61	0.64
シングルモーダル（聴覚）	1.00	0.66	0.46	1.00	0.79	0.62	0.92	0.83	0.80

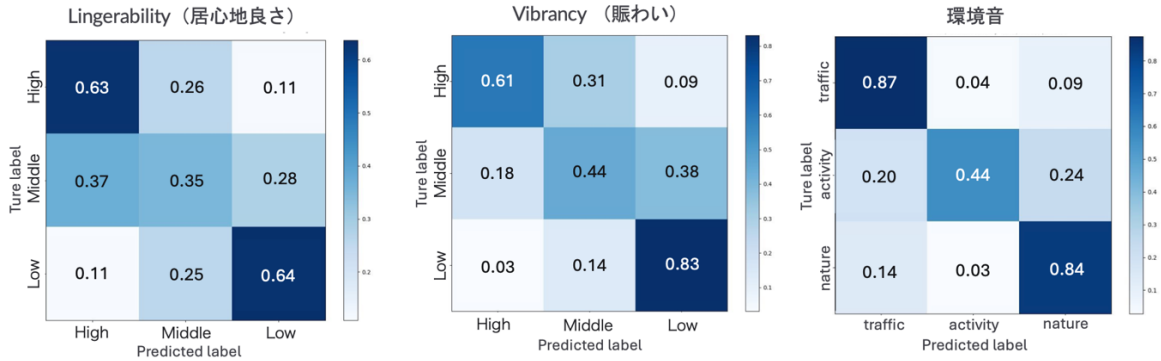

図-5 各指標に対するマルチモーダル AI の混同行列

ここでは，評価値 1〜3 に対応するクラスラベルとして 1: Low，2: Middle，3: High とした．

Lingerability と Vibrancy の評価では，「High」および「Low」クラスでは比較的高い正解率が得られている．たとえば，Lingerability においては High が 0.63，Low が 0.64 であり，明確な印象として識別されやすい傾向がうかがえる．これに対し，「Middle」クラスでは正解率が相対的に低く，他クラスとの分類結果が交差する傾向が見られる．この結果は，空間性能指標が主観的なスケールに基づくことに加え，中間的なカテゴリにおいては特にクラス間の境界が曖昧になりやすいことを反映していると推察できる．また，視覚および聴覚に基づく印象は，連続的かつ重層的な性質を有していることから，その特性が混同行列にも表れていると解釈できる．

一方，環境音の分類に関しては，交通音および自然音において，それぞれ 0.87，0.84 と高い分類精度が得られている．対して，活動音は正解率が 0.44 にとどまり，分類精度が相対的に低い．これは，活動音が交通音や風切り音など，他の背景音と混在して発生することが多いため，分類が困難となる傾向を反映していると考えられる．

(2)　環境音が空間性能に与える影響分析

環境音と空間性能（Lingerability）との関係を**図-6** に示す．横軸に LI スコアを，縦軸には各区間内における音環境ラベル（交通音・活動音・自然音）のクラス内割合を示している．クラス内割合とは，各環境音ラベルに分類されたデータのうち，LI スコアが各スコア区間に分布する割合を示している．

LI が低い区間（0.0〜0.3 程度）では交通音が高い割合を占めており，空間の居心地の良さを低下させる要因となっていることが示唆される．一方，LI が高い区間（0.7〜1.0）では，自然音や活動音の割合が相対的に高く，これらの音環境が居心地の良さと正の関係を有すると考えられる．

特に自然音はスコアの上昇に伴って増加しており，視覚的な緑地要素と併せて，快適性を高める要素として機能している可能性がある．また活動音については，LI が 0.9 以上の区間で特に高い割合を示しており，居心地良さが高く評価される空間においては，人の話し声や笑い声，足音といった活動的な音が多く含まれている傾向が見て取れる．こうした活動音は，人中心の街路空間や歩行者優先空間において多く観測される傾向があることを示唆している．

以上より，交通音が移動空間の快適性に負の影響を与える一方で，自然音や活動音が快適性に寄与する可能性を示しており，都市空間における音環境の質が心理的印象形成において重要な役割を果たしていることがわかる．

(3)　視聴覚情報の相互作用における感度分析

歩行者中心空間（PC）と車両中心空間（AC）に対して，環境音を置換した際の空間性能評価への影響を**表-4** に示す．3 つの視聴覚シナリオ 1) PC+交通騒音，2) AC+騒音低減，3) AC+賑わい創出，それぞれについて，元の環境音を用いた場合と置換後の環境音を用いた場合の LI および VI の平均値を算出し，t 検定により統計的な有意差を検証した．

1) 「PC+交通騒音」シナリオにおいて，元の環境音では LI が 0.82 であるのに対し，交通騒音への置換後は評価値が 0.52 まで有意に低下し，統計的に有意な差が確認された（$p<0.01$）．同様に，VI も 0.58 から 0.41 へと有意に低下した（$p<0.01$）．

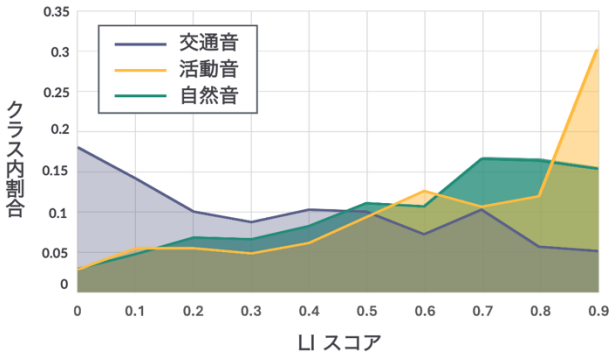

図-6　環境音ラベルごとの LI スコア分布

表-4　都市環境音の置換前後における空間性能評価の比較

	PC+交通騒音（交通騒音暴露）		AC+騒音低減（交通静穏化）		AC+都市公園の環境音（歩車共存）	
環境音	実環境音	交通騒音	実環境音	騒音低減	実環境音	都市公園
LI 平均値	0.82	0.52	0.27	0.68	0.27	0.53
VI 平均値	0.58	0.41	0.20	0.12	0.20	0.41
t 値（LI）	11.04**		-13.63**		-8.12**	
t 値（VI）	3.62**		2.54*		-7.34**	

*p<0.05, **p<0.01（各シナリオ n=100）

2)「AC+騒音低減」シナリオにおいて，騒音低減環境音への置換により LI が 0.27 から 0.68 へと有意に上昇した（$p<0.01$）．一方，VI は 0.20 から 0.12 へと有意に低下した（$p<0.05$）．

3)「AC+都市公園の環境音」シナリオにおいては，都市公園の環境音への置換により，LI が 0.27 から 0.53 へ（$p<0.01$），VI が 0.20 から 0.41 へ（$p<0.01$）といずれも有意に上昇した．

これらの結果から，視覚情報が同一であっても，環境音の違いにより空間性能評価が大きく変動することが明らかとなった．1 つ目のシナリオにおいては，元来居心地が良いと評価される歩行者中心空間(PC)に交通騒音を付与することで，居心地良さが 0.3 ポイントも低下した．この変化量は 3 つのシナリオの中で最も大きく，交通騒音の暴露が歩行者空間の快適性を著しく損なうことを定量的に示している．また，賑わいの評価も低下しており，交通騒音が空間の活気に対してもネガティブな影響を及ぼすことが示唆される．

2 つ目のシナリオでは，車両中心空間(AC)に騒音低減環境音を付与することで，居心地良さが 0.41 ポイント上昇した．これは，交通静穏化施策により交通騒音が抑制されることで，車道空間であっても居心地の良さが大幅に向上する可能性を示している．一方，賑わいの評価は若干低下したが，これは騒音低減に伴う静謐性の向上により，活動的な雰囲気が相対的に抑えられた結果であると考えられる．

3 つ目のシナリオでは，車両中心空間(AC)に都市公園の環境音を付与することで，居心地良さが 0.26 ポイント，賑わいが 0.21 ポイント上昇した．この結果は，歩車共存空間において，車両通行があっても静穏かつ人々の活動を促進する音環境を整備することで，居心地良さと賑わいの両立が可能であることを定量的に示している．特に，交通静穏化シナリオと異なり VI も向上している点は，都市公園的な音環境が人の活動音や自然音を含み，空間の活気を感じさせる要素となっていることを示唆する．

以上より，環境音は空間性能評価に対して明確に影響をあたえ，視聴覚情報の組み合わせによって評価が大きく変化することが確認された．

5. 結論

本研究では，都市空間に対する印象を視聴覚情報から推定し評価するマルチモーダル AI を開発し，移動空間の性能評価への適用可能性を検証した．具体的には，72 カ国から収集した 15,523 セグメントの歩行映像に主観評価を施し，視聴覚情報を統合的に処理するためのファインチューニングモデルを構築した．モデルの性能評価の結果，視聴覚を統合したマルチモーダルモデルは，視覚または聴覚単独モデルを上回る推定精度を達成した．Lingerability では正解率 0.56，Vibrancy では 0.67，環境音分類では 0.82 の精度を示し，視聴覚情報を統合的に処理する意義を示した．環境音と空間性能の関係分析からは，LI スコアが低い区間では交通音が支配的である一方，高い区間では自然音や活動音の割合が増加することが明らかになった．さらに，感度分析を通じて，視覚情報が同一であっても環境音の違いにより空間性能評価が大きく変動することが確認された．特に，歩行者中心空間への交通騒音付与により LI が 0.3 ポイント低下する一方，車両中心空間への交通静穏化により LI が 0.41 ポイント上昇するなど，音環境が空間評価に及ぼす影響を確認した．

本研究は，グローバルスケールでの視聴覚データセットの構築を通じて，多様な都市環境における空間性能評価の基盤を提供するとともに，視聴覚情報を統合的に処理するマルチモーダル AI の有効性を実証した．特に，抽象的な印象指標に対して視覚情報と聴覚情報を併用することで安定かつ高精度な推定が可能となることを示すとともに，環境音が空間性能評価に与える影響を定量的に明らかにした．これらの知見は，視聴覚統合的な空間評価の理論的枠組みを提示した点に学術的意義がある．

一方で，本研究にはいくつかの限界も存在する．第一に，アノテーションの制約として，専門知識を有する評価者によるラベル付けに依拠しているため，空間の多様な利用者の視点が十分に反映されていない．クラウドソーシングやソフトラベリング等の手法を用いて，より幅広い属性の利用者による評価を収集することが今後の課題である．第二に，本研究では各セグメントを 6 秒の静的なスナップショットとして扱っているが，実際の都市空間体験は時間経過に伴う動的な変化を含むため，時系列情報を考慮したモデル拡張が必要である．第三に，本研究では歩行映像の視聴覚情報を主な入力としているが，周辺の土地利用，施設配置，交通ネットワークといった空間の構造的特性との関連性を検証する必要がある．最後に，本研究の主眼はモデルの開発であり，実際の都市空間整備計画への適用性については十分に言及できていない．これらの限界を踏まえた上で，今後の研究展開が求められる．

今後の都市空間デザインにおいては，視覚環境と聴覚環境を複合的に考慮した統合的アプローチが，

都市空間の質向上居心地が良く歩きたくなる空間の実現に不可欠である．本研究で開発したマルチモーダル AI は，視聴覚統合的な空間性能評価を実現する実践的手法として，都市の質向上に寄与することが期待される．

謝辞：本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2138 および，JST/COI-NEXT（課題番号 JPMJPF2009）の支援により実施された．ここに記して謝意を表する．

REFERENCES

1) Carrera, L.: Active aging and urban policies: the space as an instrument for an inclusive and sustainable city, Frontiers in Sociology, Vol. 8, 1257926, 2023.
2) Molinsky, J. and Forsyth, A.: Climate change, aging, and well-being: how residential setting matters, Housing Policy Debate, Vol. 33, No. 5, pp. 1029–1054, 2022.
3) Bergefurt, L., Kemperman, A., van den Berg, P., Borgers, A., van der Waerden, P., Oosterhuis, G. and Hommel, M.: Loneliness and life satisfaction explained by public-space use and mobility patterns, International Journal of Environmental Research and Public Health, Vol. 16, No. 21, 4282, 2019.
4) Lorenzo, M., Ríos-Rodríguez, M. L., Suárez, E., Hernández, B. and Rosales, C.: Quality analysis and categorisation of public space, Heliyon, Vol. 9, No. 3, e13861, 2023.
5) Răducan, R., Lozsa, J., Vîrga, D. and others: New integrative model of the quality of urban life: a systematic review, Social Indicators Research, Vol. 179, pp. 895–921, 2025.
6) Eggimann, S.: The potential of implementing superblocks for multifunctional street use in cities, Nature Sustainability, Vol. 5, pp. 406–414, 2022.
7) Moreno, C., Allam, Z., Chabaud, D., Gall, C. and Pratlong, F.: Introducing the “15-minute city”: sustainability, resilience and place identity in future post-pandemic cities, Smart Cities, Vol. 4, pp. 93–111, 2021.
8) Allam, Z., Bibri, S. E., Jones, D. S., Chabaud, D. and Moreno, C.: Unpacking the '15-minute city' via 6G, IoT, and digital twins: towards a new narrative for increasing urban efficiency, resilience, and sustainability, Sensors, Vol. 22, No. 4, 1369, 2022.
9) Thornton, L. E., Schroers, R. D., Lamb, K. E. and others: Operationalising the 20-minute neighbourhood, International Journal of Behavioral Nutrition and Physical Activity, Vol. 19, 15, 2022.
10) 浅見 泰司, ウォーカブル推進都市, 日本不動産学会誌, 33 巻, 3 号, p. 54-58, 2019 [Asami, Y.: Walkability promotion city, The Japanese Journal of Real Estate Sciences, Vol. 33, No. 3, pp. 54–58, 2019.]
11) Landis, B. W., Vattikuti, V. R., Ottenberg, R. M., McLeod, D. S. and Guttenplan, M.: Modeling the roadside walking environment: pedestrian level of service, Transportation Research Record, Vol. 1773, No. 1, pp. 82–88, 2001.
12) Fruin, J. J.: Designing for pedestrians: a level-of-service concept, Highway Research Record, 1971.
13) Petritsch, T. A., Landis, B. W., Huang, H. F. and Dowling, R.: Pedestrian level-of-service model for arterials, Transportation Research Record, Vol. 2073, No. 1, pp. 58–68, 2008.
14) Burton, L. M., Matthews, S. A., Leung, M., Kemp, S. P. and Takeuchi, D. T. (eds.): Communities, neighbourhoods, and health: expanding the boundaries of place, Springer, 2011.
15) Keizer, K. and others: The spreading of disorder, Science, Vol. 322, pp. 1681–1685, 2008.
16) Miralles-Guasch, C., Dopico, J., Delclòs-Alió, X., Knobel, P., Marquet, O., Maneja-Zaragoza, R., Schipperijn, J. and Vich, G.: Natural landscape, infrastructure, and health: the physical activity implications of urban green space composition among the elderly, International Journal of Environmental Research and Public Health, Vol. 16, 3986, 2019.
17) Lynch, K.: The image of the city, MIT Press, 1960.
18) Dai, T. and Zheng, X.: Understanding how multi-sensory spatial experience influences atmosphere, affective city image and behavioural intention, Environmental Impact Assessment Review, Vol. 89, 106595, 2021.
19) Spence, C.: Senses of place: architectural design for the multisensory mind, Cognitive Research, Vol. 5, 46, 2020.
20) Gu, Y. and others: Designing effective image-based surveys for urban visual perception, Landscape and Urban Planning, Vol. 260, 105368, 2025.
21) Naik, N., Philipoom, J., Raskar, R. and Hidalgo, C.: Streetscore – predicting the perceived safety of one million streetscapes, 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 793–799, 2014.
22) Quercia, D., O'Hare, N. K. and Cramer, H.: Aesthetic capital: what makes London look beautiful, quiet, and happy?, Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14), pp. 945–955, Association for Computing Machinery, New York, 2014.
23) Salesses, P., Schechtner, K. and Hidalgo, C. A.: The collaborative image of the city: mapping the inequality of urban perception, PLOS ONE, Vol. 8, No. 7, e68400, 2013.
24) Dubey, A., Naik, N., Parikh, D., Raskar, R. and Hidalgo, C. A.: Deep learning the city: quantifying urban perception at a global scale, in Leibe, B., Matas, J., Sebe, N. and Welling, M. (eds.), Computer Vision – ECCV 2016, Lecture Notes in Computer Science, Vol. 9905, Springer, Cham, 2016.
25) Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L. and Hidalgo, C. A.: Computer vision uncovers predictors of physical urban change, Proceedings of the National Academy of Sciences of the United States of America, Vol. 114, No. 29, pp. 7571–7576, 2017.
26) Rossetti, T., Lobel, H., Rocco, V. and Hurtubia, R.: Explaining subjective perceptions of public spaces as a function of the built environment: a massive data approach, Landscape and Urban Planning, Vol. 181, pp. 169–178, 2019.
27) Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H. and Ratti, C.: Measuring human perceptions of a large-scale urban region using machine learning, Landscape and Urban Planning, Vol. 180, pp. 148–160, 2018.

28) Wei, J., Yue, W., Li, M. and Gao, J.: Mapping human perception of urban landscape from street-view images: a deep-learning approach, International Journal of Applied Earth Observation and Geoinformation, Vol. 112, 102886, 2022.
29) Wang, L., Han, X., He, J. and Jung, T.: Measuring residents' perceptions of city streets to inform better street planning through deep learning and space syntax, ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 190, pp. 215–230, 2022.
30) Kang, Y., Abraham, J., Ceccato, V., Duarte, F., Gao, S., Ljungqvist, L., Zhang, F., Näsman, P. and Ratti, C.: Assessing differences in safety perceptions using GeoAI and survey across neighbourhoods in Stockholm, Sweden, Landscape and Urban Planning, Vol. 236, 104768, 2023.
31) Ogawa, Y., Oki, T., Zhao, C., Sekimoto, Y. and Shimizu, C.: Evaluating the subjective perceptions of streetscapes using street-view images, Landscape and Urban Planning, Vol. 247, 105073, 2024.
32) 木崎 礼雄, 柳沼 秀樹, 大山 雄己, 寺部 慎太郎, 鈴木 雄, 街路景観を考慮した歩行者経路選択モデルに基づく街路空間整備評価, 土木学会論文集, 80 巻, 20 号, 2024. [Kizaki, R., Yaginuma, H., Oyama, Y., Terabe, S. and Suzuki, Y.: Dynamic pedestrian route choice model considering street landscape and spatial characteristics, Japanese Journal of JSCE, Vol. 80, No. 20, 2024.]
33) Ren, X., Wei, P., Wang, Q., Sun, W., Yuan, M., Shao, S., Zhu, D. and Xue, Y.: The effects of audio-visual perceptual characteristics on environmental health of pedestrian streets with traffic noise: a case study in Dalian, China, Frontiers in Psychology, Vol. 14, 1122639, 2023.
34) McAlexander, T. P., Gershon, R. R. and Neitzel, R. L.: Street-level noise in an urban setting: assessment and contribution to personal exposure, Environmental Health, Vol. 14, 18, 2015.
35) Axelsson, Ö., Nilsson, M. E. and Berglund, B.: A principal components model of soundscape perception, Journal of the Acoustical Society of America, Vol. 128, No. 5, pp. 2836–2846, 2010.
36) Kang, J. and Zhang, M.: Semantic differential analysis of the soundscape in urban open public spaces, Building and Environment, Vol. 45, No. 1, pp. 150–157, 2010.
37) Jo, H. I. and Jeon, J. Y.: Urban soundscape categorization based on individual recognition, perception, and assessment of sound environments, Landscape and Urban Planning, Vol. 216, 104241, 2021.
38) Liu, J., Kang, J., Luo, T., Behm, H. and Coppack, T.: Spatiotemporal variability of soundscapes in a multiple functional urban area, Landscape and Urban Planning, Vol. 115, pp. 1–9, 2013.
39) Kang, J., Aletta, F., Gjestland, T. T., Brown, L. A., Botteldooren, D., Schulte-Fortkamp, B., Lercher, P., van Kamp, I., Genuit, K., Fiebig, A., Bento Coelho, J. L., Maffei, L. and Lavia, L.: Ten questions on the soundscapes of the built environment, Building and Environment, Vol. 108, pp. 284–294, 2016.
40) Piczak, K. J.: Environmental sound classification with convolutional neural networks, 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6, 2015.
41) Lezhenin, I., Bogach, N. and Pyshkin, E.: Urban sound classification using long short-term memory neural network, 2019 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 57–60, 2019.
42) Bubashait, M. and Hewahi, N.: Urban sound classification using DNN, CNN & LSTM: a comparative approach, 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pp. 46–50, 2021.
43) Verma, D., Jana, A. and Ramamritham, K.: Classification and mapping of sound sources in local urban streets through AudioSet data and Bayesian optimized neural networks, Noise Mapping, Vol. 6, No. 1, pp. 52–71, 2019.
44) Hou, Y., Ren, Q., Zhang, H., Mitchell, A., Aletta, F., Kang, J. and Botteldooren, D.: AI-based soundscape analysis: jointly identifying sound sources and predicting annoyance, Journal of the Acoustical Society of America, Vol. 154, No. 5, pp. 3145–3157, 2023.
45) Zhao, T., Liang, X., Tu, W., Huang, Z. and Biljecki, F.: Sensing urban soundscapes from street view imagery, Computers, Environment and Urban Systems, Vol. 99, 101915, 2023.
46) Rui, Q., Gu, K. and Cheng, H.: Integrating street-view images to quantify the urban soundscape: case study of Fuzhou City's main urban area, Journal of the Acoustical Society of America, Vol. 156, No. 4, pp. 2090–2105, 2024.
47) Hall, E. T.: The hidden dimension, Anchor, 1966.
48) Jeon, J. Y. and Jo, H. I.: Effects of audio-visual interactions on soundscape and landscape perception and their influence on satisfaction with the urban environment, Building and Environment, Vol. 169, 106544, 2020.
49) Verma, D., Jana, A. and Ramamritham, K.: Machine-based understanding of manually collected visual and auditory datasets for urban perception studies, Landscape and Urban Planning, Vol. 190, 103604, 2019..
50) Verma, D., Jana, A. and Ramamritham, K.: Predicting human perception of the urban environment in a spatiotemporal urban setting using locally acquired street view images and audio clips, Building and Environment, Vol. 186, 107340, 2020.
51) Chen, P. and others: Cross-modal urban sensing: evaluating sound-vision alignment across street-level and aerial imagery, arXiv, abs/2506.03388, 2025.
52) Chen, S., Chen, Z., Hong, J., Zhuang, X., Su, C. and Ding, Z.: Exploring the relationship between audio–visual perception in Fuzhou universities and college students' attention restoration quality using machine learning, Frontiers in Psychology, Vol. 16, 1572426, 2025.
53) Aumond, P., Can, A., De Coensel, B., Ribeiro, C., Botteldooren, D. and Lavandier, C.: Global and continuous pleasantness estimation of the soundscape perceived during walking trips through urban environments, Applied Sciences, Vol. 7, 144, 2017.
54) Chu, S., Narayanan, S. and Kuo, C.-C. J.: Environmental sound recognition with time–frequency audio features, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, No. 6, pp. 1142–1158, 2009.
55) Day, G., and Gwilliam, J.:Living Architecture, Living Cities: Soul-Nourishing Sustainability, Routledge, 2019.
56) 中村文彦, 国際交通安全学会 都市の文化的創造的機

能を支える公共交通のあり方研究会:余韻都市 ニューローカルと公共交通, 鹿島出版会, 2022. [Nakamura, F. and the research group member on public transport supporting cultural-and-creative function of cities by IATSS: Afterglow Cites –New Local and Public Transport, Kajima Institute Publishing, 2022]
57) Tan, J. K. A., Hasegawa, Y. and Lau, S.-K.: A comprehensive environmental sound categorization scheme of an urban city, Applied Acoustics, Vol. 199, 109018, 2022.
58) Jeon, J. Y. and Hong, J. Y.: Classification of urban park soundscapes through perceptions of the acoustical environments, Landscape and Urban Planning, Vol. 141, pp. 100–111, 2015.
59) Stevens, S. S., Volkmann, J. and Newman, E. B.: A scale for the measurement of the psychological magnitude pitch, Journal of the Acoustical Society of America, Vol. 8, No. 3, pp. 185–190, 1937.
60) Dossou, B. F. P. and Gbenou, Y. K. S.: FSER: deep convolutional neural networks for speech emotion recognition, 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3526–3531, 2021.
61) Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations (ICLR 2015), pp.1–14.
62) Li, Z. and Hoiem, D.: Learning without forgetting, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, No. 12, pp. 2935–2947, 2017.
63) Law, S. et al.: Street-Frontage-Net: urban image classification using deep convolutional neural networks, International Journal of Geographical Information Science, Vol. 34, No. 4, pp. 681–707, 2020.

(Received ?? ??, 2025)
(Accepted ?? ??, 2025)

AI-BASED AUDIOVISUAL EVALUATION FOR FOSTERING HUMAN-CENTERED URBAN DESIGN

Kanyou SOU, Kento YOH, Kenji DOI and Yuta NAKASHIMA

In urban space design that requires a shift from automobile-oriented to human-centered approaches, it is essential to redefine mobility spaces from mere thoroughfares to places that support lingering and diverse activities, and to establish a theoretical framework that reflects user perceptions shaped by the urban environment in planning and design. In this study, we annotated approximately 15,000 segments extracted from walking videos across 72 countries based on subjective evaluations, and constructed a multimodal AI model that integrates spatial attributes such as visual and auditory environments for comprehensive processing. As a result, the model integrating audiovisual information achieved higher prediction accuracy compared to models using visual or auditory information alone. By analyzing the distribution of spatial performance evaluations for each environmental sound category from the model outputs, we clarified the relationship between them, and demonstrated the validity and applicability of the audiovisual-based evaluation model through sensitivity analysis.