Vol.2021-MUS-131 No.51
Vol.2021-SLP-137 No.51
2021/6/19

スタジアムにおける大規模群集の音響イベント分析

坂東 宜昭[1,a)]　大西 正輝[1]　内藤 航[1]　保高 徹生[1]

概要：本稿では，転移学習を活用したスタジアムでの群衆の音響イベント分析について述べる．当面は新型コロナウイルスと共存しなければならない可能性が高い現在，安全な生活様式を確立するために，人々の行動を定量的に評価する技術が不可欠である．本研究では，大勢の観客が参加するスタジアムでの行動分析を目的として，音響イベント検出システムを開発する．サッカーなどのスタジアムでは，声出し応援の禁止・拍手による応援の導入を含む様々な感染症対策が実施されている．音響イベント検出技術を活用すれば，これらの遵守状況を定量的に確認でき，感染リスクの評価に役立つ．このような検出システムを学習するには，ラベル付きデータが不足する課題があるが，AudioSet などの大規模データセットでの事前学習モデルを転移学習することで克服する．実験では，最新の学習済みモデルの本タスクにおける性能を比較評価し，未知環境への頑健性や学習データの準備コストの観点から各モデルの実用しやすさを報告する．

1. はじめに

新型コロナウイルス（COVID-19）の完全な収束が難しい現在では，当面は共存しながら生活しなければならない可能性が高い．このような with コロナ社会における安全な生活様式を確立するには，人々の行動を定量的に評価する技術が不可欠である．本ウイルスの主な感染経路は，感染者の会話や咳・くしゃみなどで出た飛沫を直接吸入する飛沫感染経路と，広範に拡散した飛沫核を吸入する飛沫核感染経路，飛沫（核）が付着した物体に触れた手を介する接触感染経路であることが知られている [1]．これらの感染経路を抑制できるような生活様式を確立できれば，より安全な社会生活が迎えられると考えられる．

本研究では，大規模に開催されるイベントでの新型コロナウイルス感染症のリスク評価を目的として，群衆の発する音響イベントを定量化する．大勢の人が行き交う空間では，感染者と直接接触する機会は少なく，会話や咳・くしゃみといった飛沫（核）を発生させる行為の頻度を定量化することが重要である．これらの検出は，従来音響イベント分析 [2–4] として広く研究されており，実環境で頑健に動作するシステムを構築できれば，広い範囲を効率的に計測できるようになる．このような行為の頻度は，感染リスク評価モデル [1,5] のパラメータの 1 つにもなっており，感染予防策が効果的に実施されているかの評価に役立つ．

音響イベントの分類 [2–4] や検出 [6–9] では，深層学習を用いた枠組み [2–4,6,8,9] が高い性能を達成しているが，

図 1　スタジアムに設置した音響センサ

音響イベントの正解ラベルが付与された教師付き学習データセットの構築コストが課題であった．特に，マスク着用が徹底され不要な声出しが自粛されている現在では，コロナウイルス感染症が蔓延する以前の音環境とは大きく変化している場合も多く，深層学習に必要な大規模な学習データセットの準備は難しい．

小規模データセットでも効率的に音響イベント検出を学習するための 1 つの枠組みとして，転移学習 [2,10–13] が注目されている．AudioSet [3] のように，5000 時間を超える大規模な（弱）ラベル付きデータセットが公開されており，このようなデータセットで学習したニューラルネットワークのパラメータが公開されている [2,14,15]．学習済みモデルを目的タスクのデータセットで再学習することで，データセットの不足を克服しながら高い性能を実現できる．しかし多くの論文では，公開データセットにおける転移学習の性能評価は報告されているが，実運用タスクでの

[1] 産業技術総合研究所
[a)] y.bando@aist.go.jp

図 2 音響センサの構成

有効性の評価や議論の報告は少ない．

本稿では，転移学習を活用した大規模イベントでの群衆の音響イベント分析について報告する．具体的には，サッカースタジアムでの音響イベント検出システム（図 1）について述べる．日本プロサッカーリーグ（J リーグ）の公式戦では，観客動員率を 50%以下に抑えた過密環境の回避，アルコール消毒の設置・啓蒙だけでなく，声出し応援の禁止・拍手による応援など，新しい応援スタイルも導入・啓蒙して感染症対策が実施されている．このようなスタジアムでの感染症対策の遵守状況を定量的に確認するため，観客の発する音響イベントを検出するシステムを構築した．特に，最新の学習済みモデルの本タスクにおける転移学習の性能を比較評価し，未知環境での頑健性や学習データの準備コストの観点から各モデルの実用しやすさを評価した．

2. 大規模群衆の音響イベント検出システム

本節では，サッカースタジアムでの計測実験に用いた音響イベント検出システムについて説明する．

2.1 スタジアムの特性

サッカースタジアムでは一般に，サッカーフィールドを囲むように最大数万程度の客席がすり鉢状に設けられている．また，サッカーの試合は 2 チームが勝敗を競うため，これらの座席はそれぞれのサポータが分離されるように区分けされている．これらの特性から，観客の応援スタイルはスタジアム全体で一様ではなく，区画によって異なっている可能性がある．そのため，1 箇所での定点観測ではなく，スタジアム内に複数のセンサを分散配置して計測する必要がある．また，多くのスタジアムは屋外会場となっているため，電源の確保が難しい．

2.2 音響センサ

図 2 に実験で用いた音響センサの構成示す．本システムでは，観測混合音から音響イベントの種類だけでなく，その方向・空間的な広がり推定できるように，マイクロホンアレイ（マイクアレイ）を採用した．本マイクアレイは，株式会社システムインフロンティアの RASP-ZX を用いて構築しており，16 kHz・24 bit で 16 チャネル同期信号を収録できる．ただし，本稿では収録信号のうち 1 チャネルのみを用いたモノラル音響イベント検出システムに注目し，多チャネル拡張については今後の課題とする．観測した音響信号は，シングルボードコンピュータを用いてタイムスタンプ付きで収録する．リアルタイムの音響イベント分析にも対応するため，シングルボードコンピュータには組み込み GPGPU（General-purpose computing on graphics processing units）ボードである NVIDIA Corp. の Jetson Xavier NX を用いた．本システムはバッテリー駆動でき，予備実験では 10 時間程度の多チャネル信号を収録できた．

2.3 音響イベント検出

本システムでは，学習済みの音響イベント分類モデルを転移学習して短時間のクリップに対する音響イベント識別器を構築し，本識別器を時間方向にスライドさせて音響イベントを検出する．転移学習には，AudioSet での学習済みモデルが公開されている，以下の 3 種を検討する:

VGGish[*1] [14]: 本モデルは VGG [16] 型の畳み込みニューラルネットワーク（CNN）で構成され，音響イベント識別の事前学習により 128 次元の埋め込みベクトル列を出力する．本稿では，得られた埋め込みベクトル列に対して，カーネルサイズ 3・チャネル数 512 の 1 次元畳み込み層と ReLU 関数を適用したのち，平均プーリングを施し出力層を経て事後確率を得る．これら 2 層のみ学習した．

OpenL3[*2] [15]: 本モデルは，AVC (audio-visual correspondence) [15, 17] と呼ばれる，動画と音響信号の特徴量抽出をそれぞれ自己教師あり学習する枠組みを用いて学習されたモデルである．AudioSet は，Google 社の動画配信サービス YouTube の動画データから生成されているため，元動画を参照することで自己教師あり学習できる．本モデルは，6144 次元の埋め込みベクトル列を出力するため，VGGish の場合と同様に，1 次元畳み込み層と出力層を追加学習して事後確率を得る．

PANN[*3] [2]: 14 層の CNN を用いて AudioSet での音響イベント識別を事前学習したモデルである．本モデルは音響クリップに対し 2048 次元の埋め込みベクトルを出力する．本稿では，出力層と共にネットワーク全体を finetune した．また，比較のため本モデルをランダム初期値から学習した場合も検討する．

これらのネットワークは，4 節にて以下の 2 つの観点から比較し，最も性能の良かったものを採用する: 1）設置場所やスタジアムの違いに起因するドメインミスマッチに対し

*1 https://github.com/tcvrick/audioset-vggish-tensorflow-to-pytorch
*2 https://github.com/marl/openl3
*3 https://github.com/qiuqiangkong/audioset_tagging_cnn

表 1 アノテーションしたラベルとアノテーション区間内での頻度

ラベル	頻度 [%]
拍手	36.8
（非意図的な）大人数歓声	3.0
少人数会話	3.0
咳・くしゃみ	0.1
楽器応援	5.8
選手掛声	19.0
ホイッスル	1.0
放送	10.0

て，どの程度汎化性能を有するか，2）学習済みモデルの転移学習にどの程度のアノテーションが必要か．

3. 計測実験

本節では，音響センサシステムを用いたサッカースタジアムでの計測実験について述べる．

3.1 スタジアムでの計測

2021 年 4 月 3 日の愛知県豊田市 豊田スタジアムと 2021 年 4 月 11 日の東京都調布市 味の素スタジアムにおいて，J リーグ及び名古屋グランパスエイト，FC 東京の協力のもと，計測実験を行った（図 1）．本実験では，スタジアムの観客席付近に分散して 9（豊田スタジアム），8（味の素スタジアム）台のマイクアレイを設置し，観客の発する音響イベントを含む環境音を収集した．収録した音響信号は，個々の音声ではなく喧騒の計測のために用い，音声認識や会話記録は行っていない．また，実験の目的・方法をスタジアム内に掲示し，観客の方々へ計測実験を周知した．

3.2 学習・性能評価のためのアノテーション

音響イベント識別器を学習するために，一部の収録信号にアノテーションを行った．サッカースタジアムでの観客の応援スタイルの分析を目的に，表 1 に示す 8 種の音イベントの開始・終了時刻を付与した．声出し応援は禁止されているが，チャンス・ピンチ等において大人数が一斉に非意図的な歓声をあげることがあるため，大人数の歓声をアノテーションしている．一方，音声であっても，隣席との会話など，数人での会話は，少人数会話として別のラベルを付与した．また，楽器応援は，公式に認められている打楽器による応援である．

アノテーションは，豊田スタジアムと味の素スタジアムで計測した音響信号のうち，それぞれ 1 台について試合開始時間から試合終了までを含む 2 時間分（それぞれ T1, A1 とする）を，また豊田スタジアムで計測した他 6 台の概ね試合開始時刻からそれぞれ 30 分の区間（それぞれ T2 から T7 とする）をアノテーションした．表 1 に示すように，アノテーションした区間内では，大人数歓声（3.0%）に比べて拍手（36.8%）の頻度が高く，主要な応援方法になっていることが分かる．また，少人数での会話は 3.0%であった．ただし，試合中のマスク着用率は平均 94%であった[*4]．

4. 音響イベント識別器の性能評価

本節では，3.2 節で述べたアノテーション付き録音信号を用いて学習した音響イベント識別器の性能を報告する．

4.1 実験設定

本評価実験では，以下のようにラベル付きデータを分割し，学習と評価を行った:

学習データ: 豊田スタジアムでの録音 T1 の前半 1.0 時間および T2 から T7 の計 4.0 時間
評価データ T: T1 のうち後半 1.0 時間
評価データ A: 味の素スタジアムでの録音 A1 2.0 時間

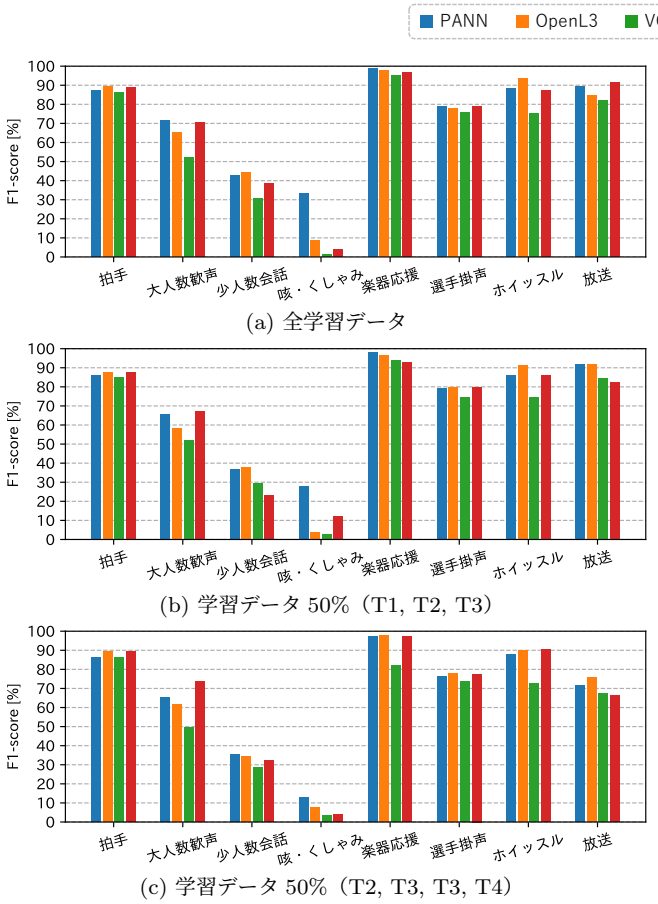
学習データのうち一部のセンサのデータを欠落させることで，学習データの減少に対する頑健性と設置場所とセンサ個体の違いに対する頑健性を確認する．評価データ T と評価データ A での性能差から，スタジアム・試合の違いに対する頑健性を確認する．ただし，評価データ A には楽器応援は含まれていなかった．モデル選択に使用する検証データには，学習データからランダムに分割した 10%を用いた．

各データを 2 秒間のクリップに分割し，それぞれのクリップに対象の音響イベント（表 1）が含まれているか否かを交差エントロピー誤差を用いて学習した．本実験では，2.3 節で述べた 3 種の学習済みモデル (PANN, VGGish, OpenL3) を転移学習して比較した．また，ベースラインとして PANN と同じアーキテクチャ（CNN14）をランダム初期値から学習した場合とも比較した．最大 50 エポック学習し，検証データでの交差エントロピー誤差が最小となった結果を評価に用いた．バッチサイズは 32 クリップとした．重み更新には学習率 1.0×10^{-4} の Adam [18] を用いた．評価尺度には適合率と再現率の調和平均である F1 スコアを用いた．ただしネットワークが出力した事後確率の閾値は，検証データの F1 スコアが最大となるように 0.0 から 1.0 の範囲で 0.1 間隔で最適化して決定した．

4.2 実験結果

図 3-(a) に，学習時と同じ環境（評価データ T）における，全学習データを用いた場合の F1 スコアを示す．最も良い性能を発揮するモデルは，音響イベントの種類により異なっている．“大人数歓声”および “楽器応援”，“選手掛声”では PANN が，“拍手”および “少人数会話”，“ホイッスル” では OpenL3 が最高となり，“放送”では転移学習しない場合が最も高い F1 スコアであった．特にラベルが少

[*4] https://www.aist.go.jp/aist_j/new_research/2021/nr20210511/nr20210511.html

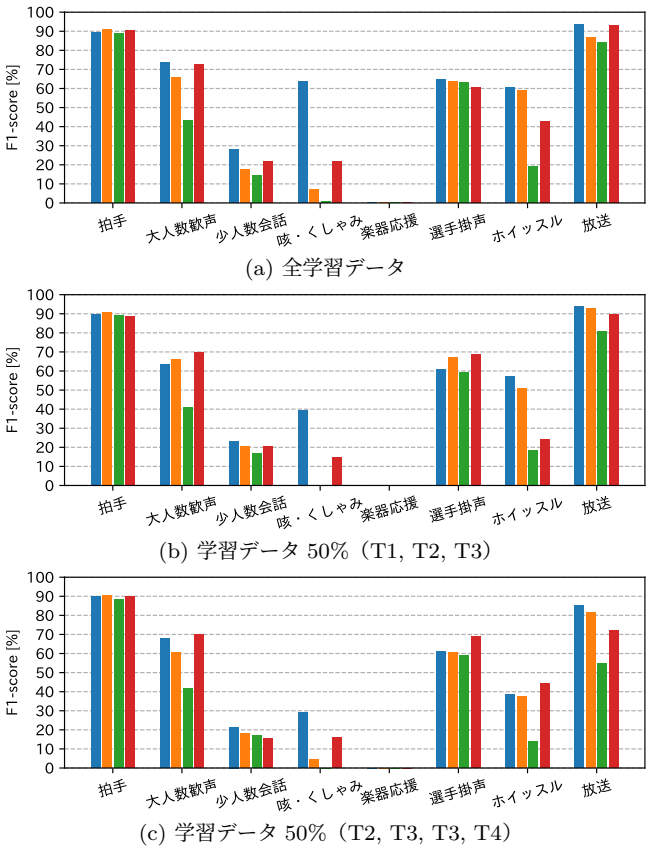


(a) 全学習データ

(b) 学習データ 50%（T1, T2, T3）

(c) 学習データ 50%（T2, T3, T3, T4）

図 3 評価データ T（同スタジアム）での音響イベント識別の性能.



(a) 全学習データ

(b) 学習データ 50%（T1, T2, T3）

(c) 学習データ 50%（T2, T3, T3, T4）

図 4 評価データ A（別スタジアム）での音響イベント識別の性能.

ない “咳・くしゃみ”では，PANN は他より 20 ポイント以上高いスコアとなった．ただし，全てのモデルが，“少人数会話”と “咳・くしゃみ”で 50%以下の性能であった．これは，会話は話者や発話内容によって大きく特徴が異なり，咳やくしゃみは全学習データのうち 0.1%程度しか含まれず汎化できなかったためと考えられる．

図 3-(b) および-(c) に，一部の学習データを欠損させた場合の検出性能を示す．図 3-(b) の通り，ラベル付き学習データを全体の 50% (2 時間) に制限しても，PANN の性能劣化は 5 ポイント程度となった．また，図 3-(c) を図 3-(b) と比較すると，評価データ T のマイクアレイ（T1）と異なるセンサ・設置場所の録音信号で学習した場合でも，“せき・くしゃみ”と “放送”を除き 5 ポイント程度の性能差であった．“放送”が 20 ポイント程度劣化した原因として，T1 の学習データのみ他より長い前半 1 時間分の録音信号を含んでおり，このデータが欠落した結果，試合実況や音楽を含む多様な放送音の識別に失敗していると考えられる．

図 4-(a) の通り，学習時と異なる環境（評価データ A）においても，“拍手”および “大人数歓声”，“放送”については，5 ポイント以内の性能劣化と高い汎化性能が得られた．“拍手”と “大人数歓声”は，どちらも多くの観客から発せられる音響イベントなので，スタジアムごとの差が小さかったと考えられる．一方，“少人数会話”および “選手掛声”，“ホイッスル”では F1 スコアが 10 ポイント以上劣化した．

“少人数会話”と “咳・くしゃみ”は飛沫を発生させる行為であるので，これらの識別性能向上は不可欠である．本問題は，既存の音声コーパス [19–21] や咳・くしゃみを含む音響イベントコーパス [22] を用いて数値混合した合成データを併用すれば，性能改善すると期待できる．

5. おわりに

本稿では，転移学習を活用したスタジアムでの群衆の音響イベント分析について報告した．4 時間程度のラベル付き学習データを最大限活用するため，学習済みモデルが公開されている PANN および OpenL3, VGGish を転移学習し，その音響イベント識別性能を評価した．これらのうち，PANN と OpenL3 が高い性能を達成し，PANN は拍手および非意図的な大人数歓声において F1 スコア 87%, 72% となった．また，転移学習しなくとも，多くの音響イベントを転移学習モデルと同程度の性能で識別できたが，頻度が 0.1%と少ない咳・くしゃみは PANN が高い性能を発揮した．さらに本モデルは，学習データが半分（2 時間）であっても性能劣化は 5 ポイントであった．また，別の設置場所の収録音 2 時間で学習しても，同じ収録場所の学習データが含まれている場合と比べ，一部のラベルを除き性能劣化は 5 ポイント程度であることを確認した．未知のス

タジアムにおいても，拍手や大人数の歓声の識別には汎化性を有し F1 スコア 70%を維持できていた．一方，少人数の会話や咳・くしゃみは，音イベントの多様さとサンプル数の少なさから，学習データと同じスタジアムの評価データで 50%以下の性能となった．

今後は，データ拡張や数値混合を活用した性能向上を進めると共に，マイクアレイを活用した音響イベントの空間分布的な定量化を目指す．大勢の歓声は，スタジアム内に広く伝播するため，設置場所の異なる複数のセンサで同じ音源を観測してしまう．マイクアレイにより得られる空間情報を活用することで，スタジアム内のどの区画から発生した音響イベントかを特定できるシステムを目指す．また，本システムで得られた歓声や会話の頻度をもとに，感染リスクの評価を進める．

6. 謝辞

本研究の一部は，日本プロサッカーリーグおよび名古屋グランパスエイト，FC 東京，豊田スタジアム，味の素スタジアムの協力のもと実施した．また，本研究の一部は国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の助成を受けた．

参考文献

[1] Jones, R. M.: Relative contributions of transmission routes for COVID-19 among healthcare personnel providing patient care, *Journal of Occupational and Environmental Hygiene*, Vol. 17, No. 9, pp. 408–415 (2020).
[2] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W. and Plumbley, M. D.: PANNs: Large-scale pretrained audio neural networks for audio pattern recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2880–2894 (2020).
[3] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M. and Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780 (2017).
[4] Kong, Q., Xu, Y., Wang, W. and Plumbley, M. D.: Audio set classification with attention model: A probabilistic perspective, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 316–320 (2018).
[5] Murakami, M., Miura, F., Kitajima, M., Fujii, K., Yasutaka, T., Iwasaki, Y., Ono, K., Shimazu, Y., Sorano, S., Okuda, T. et al.: COVID-19 risk assessment at the opening ceremony of the Tokyo 2020 Olympic Games, *Microbial risk analysis*, p. 100162 (2021).
[6] Imoto, K., Tonami, N., Koizumi, Y., Yasuda, M., Yamanishi, R. and Yamashita, Y.: Sound event detection by multitask learning of sound events and scenes with soft scene labels, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 621–625 (2020).
[7] Jin, Q., Schulam, P., Rawat, S., Burger, S., Ding, D. and Metze, F.: Event-based video retrieval using audio, *Annual Conference of the International Speech Communication Association* (2012).
[8] Kumar, A. and Raj, B.: Audio event detection using weakly labeled data, *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1038–1047 (2016).
[9] Kong, Q., Xu, Y., Wang, W. and Plumbley, M. D.: A joint detection-classification model for audio tagging of weakly labelled data, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 641–645 (2017).
[10] Zhang, S., Qin, Y., Sun, K. and Lin, Y.: Few-Shot Audio Classification with Attentional Graph Neural Networks., *INTERSPEECH*, pp. 3649–3653 (2019).
[11] Shi, B., Sun, M., Puvvada, K. C., Kao, C.-C., Matsoukas, S. and Wang, C.: Few-shot acoustic event detection via meta learning, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 76–80 (2020).
[12] Van Den Oord, A., Dieleman, S. and Schrauwen, B.: Transfer learning by supervised pre-training for audio-based music classification, *Conference of the International Society for Music Information Retrieval (ISMIR 2014)* (2014).
[13] Morgado, P., Li, Y. and Vasconcelos, N.: Learning Representations from Audio-Visual Spatial Alignment, *arXiv preprint arXiv:2011.01819* (2020).
[14] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B. et al.: CNN architectures for large-scale audio classification, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135 (2017).
[15] Cramer, J., Wu, H.-H., Salamon, J. and Bello, J. P.: Look, listen, and learn more: Design choices for deep audio embeddings, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3852–3856 (2019).
[16] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
[17] Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J. and Torralba, A.: The sound of pixels, *Proceedings of the European conference on computer vision (ECCV)*, pp. 570–586 (2018).
[18] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
[19] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, *Journal of the Acoustical Society of Japan (E)*, Vol. 20, No. 3, pp. 199–206 (1999).
[20] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (2003).
[21] Sonobe, R., Takamichi, S. and Saruwatari, H.: JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, *arXiv preprint arXiv:1711.00354* (2017).
[22] Orlandic, L., Teijeiro, T. and Atienza, D.: The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms, *arXiv preprint arXiv:2009.11644* (2020).