ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414834

TEXT-TO-AUDIO GROUNDING: BUILDING CORRESPONDENCE BETWEEN CAPTIONS AND SOUND EVENTS

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, Kai Yu

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai, China
{*wsntxxn, richman, mengyuewu, kai.yu*}@sjtu.edu.cn

ABSTRACT

Automated Audio Captioning is a cross-modal task, generating natural language descriptions to summarize the audio clips' sound events. However, grounding the actual sound events in the given audio based on its corresponding caption has not been investigated. This paper contributes an *Audio-Grounding* dataset[1], which provides the correspondence between sound events and the captions provided in Audiocaps, along with the location (timestamps) of each present sound event. Based on such, we propose the text-to-audio grounding (TAG) task, which interactively considers the relationship between audio processing and language understanding. A baseline approach is provided, resulting in an event-F1 score of 28.3% and a Polyphonic Sound Detection Score (PSDS) score of 14.7%.

***Index Terms*—** text-to-audio grounding, sound event detection, dataset, deep learning.

1. INTRODUCTION

Using natural language to summarise audio content, commonly referred as Automated Audio Captioning (AAC), has attracted much attention in recent studies [1, 2, 3, 4]. Compared with other audio processing tasks like Acoustic Scene Classification (ASC) and Sound Event Detection (SED), which aim to categorize audio into specific scenes or event labels, AAC allows the model to describe audio content in natural language, a much more unrestricted text form. AAC can thus be seen as a less structural summarization of sound events. However, the correspondence between sound event detection and natural language description is rarely investigated. To achieve human-like audio perception, a model should be able to generate an audio caption and understand natural language grounded in acoustic content, i.e., grounding (detecting) each sound event mentioned in a given audio caption to corresponding segments in that audio. Explicit grounding of sound event phrases from the corresponding audio is key to audio-oriented language understanding. Moreover, it would be beneficial for generating captions with more accurate event illustrations and localized AAC evaluation methods.

Although such an audio grounding task (text-to-audio grounding, TAG) is relatively novel in audio understanding and audio-text cross-modal research, it is related to the following problems.

Visual Grounding A similar task to TAG is object grounding in Computer Vision (CV) using images or videos. The *Flickr30k Entities* [5] is the first public dataset for image grounding. Image object grounding has become a research hotspot since then [6, 7, 8]. Recently a plethora of work focus on new datasets and approaches for video object grounding [9, 10, 11]. Like audio-text grounding, visual grounding requires a model to predict bounding boxes (2d coordinates) in an image or video frame for each object described in the caption.

Sound Event Detection (SED) SED aims to classify and localize particular sound events in an audio clip. With the growing influence of Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [12], research interest in SED has soared recently. TAG can be viewed as text-query-based SED, focusing on localizing sound events described by queries. Due to SED and TAG's intrinsic correlation, we borrow common approaches and evaluation metrics from SED as a benchmark for TAG.

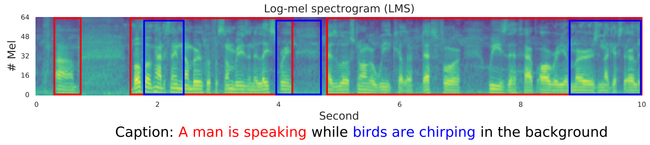

Fig. 1. An example for *TextToAudioGrounding*. For an audio clip and its corresponding caption, on- and off-set timestamps for each sound event phrase are provided. In this example, both "a man speaking" (red) and "birds chirping" (blue) point to multiple segments (presented by rectangles in the figure).

Mengyue Wu and Kai Yu are the corresponding authors.

[1] `https://github.com/wsntxxn/TextToAudioGrounding`

An audio grounding task inevitably consists of two parts. First is the extraction of sound event phrases from natural language caption, e.g., “people speak” and “dogs bark” can be obtained from the caption “people speak while dogs bark”. The second stage is concerned with traditional SED, detecting a sound event presence along with its onset and offset timestamps in the given audio clip. The prerequisite is a dataset that simultaneously provides audio, captions and the segmentation of sound events grounded from the caption. To the best of our knowledge, no existing datasets or tasks are focusing on text-to-audio grounding.

We contribute *AudioGrounding* dataset (Section 2) in this paper, providing a corresponding series of *audio - caption - sound event phrase - sound event timestamp segmentation* to enable a more interactive cross-modal research within audio processing and natural language understanding. An illustration from *AudioGrounding* is shown in Figure 1. With this dataset, we consider TAG, which localizes corresponding sound events in an audio clip from a given language description. A baseline approach for the new TAG task is also proposed, see Section 3. Section 4 details the experiment results and the analyses of such a TAG task, with conclusions provided in Section 5.

2. THE AUDIO GROUNDING DATASET

Our *AudioGrounding* dataset entails 4994 audios, with one caption per audio in the training set, five captions per audio in the validation, and test sets. We provide caption-oriented sound event tagging for each audio, along with each sound event’s segmentation timestamps. The audio sources are rooted in *AudioSet* [13] and the captions are sourced from *Audiocaps* [14].

2.1. Audio and Caption Tailoring

AudioSet is a large-scale manually-annotated sound event dataset. Each audio clip has a duration of up to ten seconds, containing at least one sound event label. *AudioSet* consists of a 527 event ontology, encompassing most everyday sounds.

Audiocaps [14] is by far the largest AAC dataset, consisting of 46,000+ audio clips ($\approx$ 127 hours) collected from *AudioSet*. One human-annotated caption is provided for the training dataset while five captions for validation and test sets, respectively. Since the entire *Audiocaps* dataset is a subset of *AudioSet*, sound event labels can be obtained for each audio clip in *Audiocaps*.

It should be noted that though *AudioSet* provides sound tags and *Audiocaps* consists of descriptive captions, there is no direct link between these two annotations. As we would like to enhance the diversity of the sound events included, we selectively choose audio clips with more than four sound tags, resulting in 4994 audio clips sourced from *Audiocaps*. For a successful text-to-audio grounding, each audio clip should have not only a caption description (“A man is speaking while birds are chirping in the background”), but also the corresponding sound event phrases retrieved from the caption (“A man is speaking”, “bird are chirping”), and the on- and off-sets of these sound events.

2.2. Annotation Process

Our annotation process is decoupled into two stages: (1) sound event phrases are extracted automatically from captions; (2) we invite annotators to merge extracted phrases that correspond to the same sound event and provide the duration segmentation of each sound event.

A. Extracting Sound Event Phrases from Captions

As mentioned above, the sound event labels provided in *AudioSet* has no correspondence with the descriptive captions in *Audiocaps*. Therefore we first extract sound event phrases from captions using NLTK [15]. A phrase refers to a contiguous chunk of words in a caption. Following standard chunking methods, we extract noun phrases (NP) and combinations of NP and verb phrases (NP + VP). As sound descriptions usually stem from objects that sound (e.g., a cat) and verbs create the sound (e.g., meow), NP and NP + VP phrases can roughly summarize all possible sound events.

B. Phrase Merging and Segmentation

Manual phrase merging is necessary since there might be repetitive and unwanted information in extracted phrases. For example, the caption in Figure 2 is chunked into three phrases: “people”, “a small crowd are speaking” and “a dog barks”. However, “people” and “a small crowd are speaking” refer to the same sound event. Based on the extracted phrases, annotators are required to label an audio clip in a two-step process:

1. Merge phrases describing the same sound event into a single set and identify the number of sound events mentioned in the audio;
2. Segment each sound event with the on- and off-set timestamps.

2.3. Data Description

Our annotation results in a new audio-text grounding dataset: *AudioGrounding*. It contains 13,985 corresponding sound event phrases and 4994 captions (*Audiocaps*). After phrase merging, there are 10,910 sound events in total. Sound events included are quite diversified, with the most frequent sound event (“a man speaks”) accounts for no more than 2% of the dataset. The sound event phrase duration distribution is shown in Figure 3. Most segments last for less than 2 s and the event phrases consist of several such short segments in a single audio clip, like speech, dog barking and cat meowing. However, a considerable proportion of events (e.g., wind,

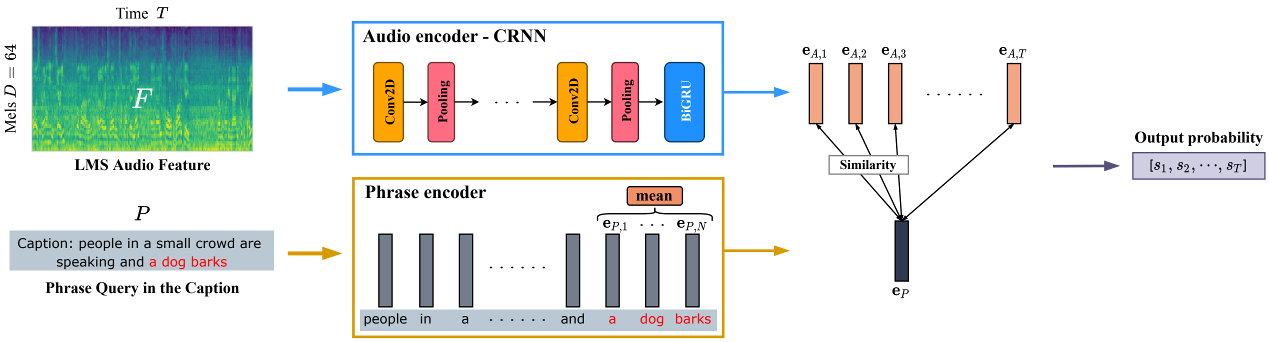
Time T
Mels $D = 64$
F
LMS Audio Feature
P
Caption: people in a small crowd are speaking and a dog barks
Phrase Query in the Caption
Audio encoder - CRNN
Conv2D
Pooling
Conv2D
Pooling
BiGRU
Phrase encoder
mean
$\mathbf{e}_{P,1}$ · · · $\mathbf{e}_{P,N}$
people in a · · · · · · and a dog barks
$\mathbf{e}_{A,1}$ $\mathbf{e}_{A,2}$ $\mathbf{e}_{A,3}$ · · · · · · $\mathbf{e}_{A,T}$
Similarity
$\mathbf{e}_P$
Output probability
$[s_1, s_2, \cdots, s_T]$

Fig. 2. The proposed baseline model structure for TAG. A CRNN encoder outputs a sequence of audio embedding $\{\mathbf{e}_{A,t}\}_{t=1}^{T}$ from the LMS input $F \in \mathbb{R}^{T \times D}$. The phrase query (containing N words) is encoded into $\mathbf{e}_P$ by taking the mean of all word embeddings $\{\mathbf{e}_{P,n}\}_{n=1}^{N}$ in the query. Prediction of the sound events' on- and off-sets are based on the similarity between $\{\mathbf{e}_{A,t}\}_{t=1}^{T}$, and $\mathbf{e}_P$.

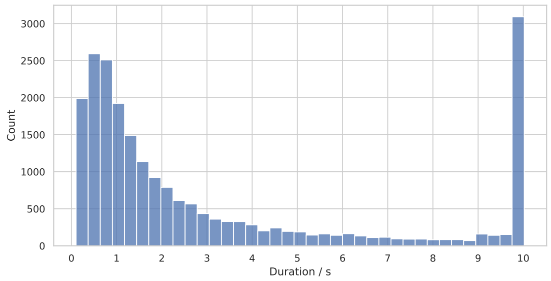

Fig. 3. Duration distribution of annotated sound events mentioned in phrases within the proposed *AudioGrounding* dataset.

train) is present in the whole clip, lasting for almost 10 s. We split the dataset according to the *Audiocaps* setting, assigning each sample to the same subset (train/val/test) in *Audiocaps*. Detailed statistics are provided in Table 1.

Table 1. Statistics of the *AudioGrounding* Dataset.

Split	#Clips	#Captions	#Sound event phrases
Train	4489	4489	12373
Val	31	155	451
Test	70	350	1161
Total	4590	4994	13958

3. TEXT-TO-AUDIO GROUNDING

Since the primary motivation regards sound event grounding from phrases in audio captions, we use two separate encoders for audio and phrase query, respectively. The input audio feature F is encoded into an embedding sequence $\{\mathbf{e}_{A,t}\}_{t=1}^{T}$ while the query encoder outputs a phrase embedding $\mathbf{e}_P$ from the phrase query P which consists of N words. Our baseline model architecture is illustrated in Figure 2. We apply $\exp(-l2)$ as the similarity metric and binary cross-entropy (BCE) loss as the training criterion, following previous work in cross-modal audio/text retrieval [16]. The similarity score between audio and phrase embedding $\mathbf{e}_{A,t}$ and $\mathbf{e}_P$ is calculated as:

$$s_t = \text{sim}(\mathbf{e}_{A,t}, \mathbf{e}_P) = \exp(-\|\mathbf{e}_{A,t} - \mathbf{e}_P\|_2) \qquad (1)$$

During training, $\mathcal{L}_{\text{BCE}}$ between an audio-phrase pair is calculated as the mean of $\mathcal{L}_{\text{BCE}}$ between $\mathbf{e}_{\text{A}}$ at each frame t and $\mathbf{e}_P$:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{T}\sum_{t=1}^{T} y_t \cdot \log(s_t) + (1 - y_t) \cdot \log(1 - s_t) \qquad (2)$$

where $y_t \in \{0, 1\}$ is a strongly labeled indicator for each t. During evaluation, $\{s_t\}_{t=1}^{T}$ is transformed to $\{\hat{y}_t\}_{t=1}^{T}, \hat{y}_t \in \{0, 1\}$ by a threshold $\phi = 0.5$, representing the presence ($\hat{y}_t = 1, s_t > \phi$) or absence ($\hat{y}_t = 0, s_t \leq \phi$) of a phrase.

Audio Encoder We adopt a convolutional recurrent neural network (CRNN) [17] as the audio encoder. The detailed CRNN architecture can be found in [18]. It consists of five convolution blocks (with padded 3 × 3 convolutions) followed by a bidirectional gated recurrent unit (BiGRU). L4-Norm subsampling layers are added between convolution blocks, reducing the temporal dimension by a factor of 4. Finally, an upsampling operation is applied to ensure the output embedding has the same sequence length as the input feature. The CRNN audio encoder outputs an embedding sequence $\{\mathbf{e}_{A,t}\}_{t=1}^{T} \in \mathbb{R}^{256}$.

Phrase Encoder For the phrase encoder, we only focus on extracting a representation for the phrase and leave out all other words in the caption. The word embedding size is also set to 256 to match $\mathbf{e}_{A,t}$. The mean of the word embeddings within a phrase is used as the representation:

$$\mathbf{e}_P = \frac{1}{N}\sum_{n=1}^{N} \mathbf{e}_{P,n} \qquad (3)$$

4. EXPERIMENTS

4.1. Experimental setup

Standard Log Mel Spectrogram (LMS) is used as the audio feature since it is commonly utilized in SED. We extract 64 dimensional LMS feature from a 40 ms window size and 20 ms window shift for each audio, resulting in $F \in \mathbb{R}^{T\times 64}$. The model is trained for at most 100 epochs using the Adam optimization algorithm with an initial learning rate of 0.001. The learning rate is reduced if the loss on the validation set does not improve for five epochs. An early stop strategy with ten epochs is adopted in the training process.

4.2. Evaluation

Since TAG shares a similar target with SED, commonly used SED metrics are adopted for TAG evaluation. Specifically, we incorporate two metrics, being event-based metrics [19] and the newly proposed polyphonic sound detection score (PSDS) [20].

- **Event-Based Metrics** (Precision, Recall, F1) attach importance to the smoothness of the predicted segments, penalizing disjoint predictions. Regarding event-F1 scores, we set a t-collar value to 100 ms (due to large amounts of short events, see Figure 3) as well as a tolerance of 20% discrepancy between the reference and prediction duration for event-based metrics.
- **PSDS** is more robust to labelling subjectivity (e.g., to create one or two ground truths for two very close dog barks) and does not depend on operating points (e.g., thresholds). The default PSDS parameters are used [20]: $\rho_{\text{DTC}} = \rho_{\text{GTC}} = 0.5, \rho_{\text{CTTC}} = 0.3, \alpha_{\text{CT}} = \alpha_{\text{ST}} = 0.0, e_{max} = 100$.

Models achieving high scores in both event-based metrics and PSDS are expected to predict smooth segments while being robust to different operating points.

Table 2. Baseline TAG performance on the *AudioGrounding* dataset. P, R, F1 represent the event-based precision, recall and, F1-score.

Model	F_1	P	R	PSDS
Random	0.04	0.02	1.56	0.00
Baseline	28.30	28.60	27.90	14.70

4.3. Results and Analyses

We present the baseline TAG performance in Table 2. The random guessing approach gives a random probability between 0 and 1 to each frame, resulting in a 0.04% event-F1 and 0.00% PSDS, indicating the difficulty of this task. In contrast, our proposed baseline model achieves 28.3% event-F1 and 14.7% PSDS, verifying its capability in audio and text understanding. Despite the significant improvement against the random approach, we find that the baseline model tends to output high probability to salient parts of an audio clip, regardless of the phrase input. An example is shown in Figure 4. The output probabilities of both phrase inputs appear to be similar in their temporal distribution. For the phrase query "young female speaking", the model assigns high presence probability to segments where either cats or female speech appear (e.g., the last two seconds). This means the model only learns prominent audio patterns but neglects the information from the phrase query. We change the phrase queries of each audio to a random phrase selected from all phrase queries of that audio. After the modification, the event-F1 is still 19.6%, indicating the insensitivity of our model to the phrase input. Further research should be conducted on the text understanding and the fusion of these two modalities.

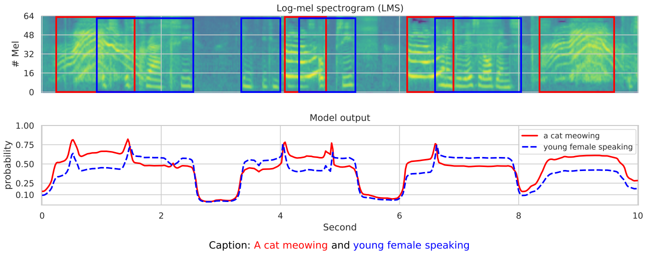

Fig. 4. An example result of a TAG prediction on the *AudioGrounding* dataset. The horizontal axis of the bottom figure denotes the output probability of a sound event according to the phrase query.

5. CONCLUSION

In this paper, we propose a Text-to-Audio Grounding task to facilitate cross-modal learning between audio and natural language further. This paper contributes an *AudioGrounding* dataset, which considers the correspondence between sound event phrases with the captions provided in Audiocaps [14] and provides the timestamps of each present sound event. A baseline approach that combines natural language and audio processing yields an event-F1 of 28.3% and a PSDS of 14.7%. We would like to explore better projection of audio and phrase embeddings as well as deeper interaction between these two modalities in future work.

6. ACKNOWLEDGEMENT

This work has been supported by National Natural Science Foundation of China (No.61901265), Shanghai Pujiang Program (No.19PJ1406300), and Startup Fund for Youngman Research at SJTU (No.19X100040009). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

7. REFERENCES

[1] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 374–378.

[2] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 830–834.

[3] Y. Koizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "The NTT DCASE2020 challenge task 6 system: Automated audio captioning with keywords and sentence length estimation," DCASE2020 Challenge, Tech. Rep., June 2020.

[4] Y. Wu, K. Chen, Z. Wang, X. Zhang, F. Nian, S. Li, and X. Shao, "Audio captioning based on transformer and pre-training for 2020 DCASE audio captioning challenge," DCASE2020 Challenge, Tech. Rep., June 2020.

[5] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2641–2649.

[6] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 817–834.

[7] R. A. Yeh, M. N. Do, and A. G. Schwing, "Unsupervised textual grounding: Linking words to image concepts," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6125–6134.

[8] H. Akbari, S. Karaman, S. Bhargava, B. Chen, C. Vondrick, and S.-F. Chang, "Multi-level multimodal common semantic space for image-phrase grounding," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 476–12 486.

[9] L. Zhou, N. Louis, and J. J. Corso, "Weakly-supervised video object grounding from text by loss weighting and object interaction," in *British Machine Vision Conference (BMVC)*, 2018, pp. 1–12.

[10] L. Chen, M. Zhai, J. He, and G. Mori, "Object grounding via iterative context reasoning," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019, pp. 1407–1415.

[11] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, "Grounded video description," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6578–6587.

[12] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, November 2018, pp. 19–23.

[13] J. F. Gemmeke, D. P. 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[14] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 119–132.

[15] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002, pp. 63–70.

[16] B. Elizalde, S. Zarar, and B. Raj, "Cross modal audio search and retrieval with joint embeddings based on text and audio," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4095–4099.

[17] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE Trans. Audio, Speech, Language Process.*, 2021.

[18] X. Xu, H. Dinkel, M. Wu, and K. Yu, "A crnn-gru based reinforcement learning approach to audio captioning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Tokyo, Japan, November 2020, pp. 225–229.

[19] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[20] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 61–65.