Overheard: Audio-based Integral Event Inference

HONGHUI XU, Kennesaw State University, USA
ZHIPENG CAI, Georgia State University, USA
LIRAN MA, Miami University, USA
YINGSHU LI, Georgia State University, USA
DAEHEE SEO, Sangmyung University, South Korea
WEI LI, Georgia State University, USA

There is no doubt that the popularity of smart devices and the development of deep learning models bring individuals too much convenience. However, some rancorous attackers can also implement unexpected privacy inferences on sensed data from smart devices via advanced deep-learning tools. Nonetheless, up to now, no work has investigated the possibility of riskier overheard, referring to inferring an integral event about humans by analyzing polyphonic audios. To this end, we propose an Audio-based integraL evenT infERence (ALTER) model and two upgraded models (ALTER-p and ALTER-pp) to achieve the integral event inference. Specifically, ALTER applies a link-like multi-label inference scheme to consider the short-term co-occurrence dependency among multiple labels for the event inference. Moreover, ALTER-p uses a newly designed attention mechanism, which fully exploits audio information and the importance of all data points, to mitigate information loss in audio data feature learning for the event inference performance improvement. Furthermore, ALTER-pp takes into account the long-term co-occurrence dependency among labels to infer an event with more diverse elements, where another devised attention mechanism is utilized to conduct a graph-like multi-label inference. Finally, extensive real-data experiments demonstrate that our models are effective in integral event inference and also outperform the state-of-the-art models.

CCS Concepts: • **Computing methodologies** → **Neural networks**; • **Information systems** → **Multimedia information systems**; • **Security and privacy**;

Additional Key Words and Phrases: Multi-label Image Recognition, Differential Privacy, Robustness

ACM Reference Format:
Honghui Xu, Zhipeng Cai, Liran Ma, Yingshu Li, Daehee Seo, and Wei Li. 2024. Overheard: Audio-based Integral Event Inference. *ACM J. Data Inform. Quality* 1, 1 (August 2024), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Nowadays, the recordings of visual and audio data capturing various scenes of people's daily life can be acquired and collected anywhere and anytime through cameras and microphones on ubiquitous smart devices [27, 28, 34]. In the meantime, with the advent of the deep learning era,

Authors' addresses: Honghui Xu, hxu10@kennesaw.edu, Kennesaw State University, 1100 South Marietta Pkwy, Marietta, Georgia, USA, 30060; Zhipeng Cai, zcai@gsu.edu, Georgia State University, P.O. Box 3965, Atlanta, Georgia, USA, 30302; Liran Ma, mal18@miamioh.edu, Miami University, 501 E High St, Oxford, Ohio, USA, 45056; Yingshu Li, yili@gsu.edu, Georgia State University, P.O. Box 3965, Atlanta, Georgia, USA, 30302; Daehee Seo, daehseo@smu.ac.kr, Sangmyung University, 20 Hongjimun 2-gil, Seoul, South Korea, 03016; Wei Li, wli28@gsu.edu, Georgia State University, P.O. Box 3965, Atlanta, Georgia, USA, 30302-3965.

1936-1955/2024/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

visual and audio data can be analyzed more effectively for providing individuals with more accurate customized services. However, the evolution of technology is a double-edged sword – such data can also be malevolently used by attackers to infer individuals' sensitive information [5, 11, 20], causing severe privacy leakage and economic loss.

So far, many works have been proposed to investigate visual and audio data-oriented privacy inference models. These visual-based approaches can successfully achieve the identification of individuals [18], the inference of individuals' activities [22], and the recognition of individuals' locations [13]. Nevertheless, these models suffer a lot of performance loss because of the poor image quality and may even become infeasible due to the constraint of camera coverage. Considering the omnidirectional coverage and easier deployment of audio sensors, some researchers change their targets to study imperceptible privacy inference attacks on audio data. These audio-based privacy inference models can be broadly classified into three categories. (i) Audio-based person identification approaches are designed by discriminating the timbres of different people [9, 23]; (ii) Sound prediction models have been developed to classify different activities' sounds of human for human activity detection [1, 4]; (iii) Environmental scene recognition schemes are devised to infer indoor and outdoor environments where human locate through distinguishing the various environmental audios [6, 31]. But, the existing works are only able to infer one specific element of an event about human, such as who they are, what they do, or where they are. Although these one-element prediction approaches can be combined to perform integral event inference, such a method lacks scalability in reality as the number of elements needs to be known or determined before model combination. What's worse, this event inference model built in a simple combination way will become more and more complicated with the increase of elements, greatly increasing implementation cost. Therefore, it is still challenging to design an effective and scalable audio-based integral event inference model.

To fill this blank, we present an Audio-based integraL evenT infERence (ALTER) model that is composed of three main components, including data preprocessing, sequential data feature learning, and multi-label inference. Our ALTER model can successfully achieve the goal of integral event inference by simultaneously leveraging the temporal correlation in the time-series audio data and the short-term co-occurrence dependency among multiple labels. Additionally, to alleviate the information loss in the sequential data feature learning, we improve ALTER model to the ALTER-p model by designing a new attention mechanism, in which we entirely exploit the audio information and the importance of all data points to get the output data features. Besides, for the purpose of inferring a sophisticated event with more various elements, the ALTER-p model is further upgraded to the ALTER-pp model, where we devise another new attention mechanism to help represent the long-term co-occurrence dependency among labels. Finally, the effectiveness of the three proposed models is evaluated and compared by conducting comprehensive real-data experiments. The multifold contributions of our work are concluded below.

- To the best of our knowledge, this is the first work to investigate an audio-based integral event inference task.
- We design ALTER, ALTER-p, and ALTER-pp models to perform the audio-based integral event inference with considering different application requirements and data characteristics.
- In our models, one novel attention mechanism is developed to retain information as much as possible in audio data feature learning, and another creative attention mechanism is implemented to capture the long-term co-occurrence dependency among multiple labels.
- We also propose a link-like multi-label inference scheme and a graph-like multi-label inference method to realize the event inference based on the short-term co-occurrence dependency and the long-term co-occurrence dependency among labels, respectively.

- Extensive real-data experiments are well conducted to validate the effectiveness of our proposed models on integral event inference and to illustrate their superiority over state-of-the-art approaches.

The rest of this paper is organized as follows. The related works are briefly summarized in Section 2. We detail our methodology in Section 3, and then conduct real-data experiments and analyze the experimental results in Section 4. After that, we propose some discussions and future works in Section 5. Finally, we end up with a conclusion in Section 6.

2 RELATED WORKS

In this section, we summarize the related works on visual-based and audio-based privacy inference models.

2.1 Visual-based Privacy Inference

With the impressive growth of deep learning in computer vision [16], attackers can maliciously detect, extract, and retrieve individuals' sensitive information in visual data via deep learning models. When one person's visual data is public on social platforms, attackers can leverage deep learning tools to automatically steal his/her private information, including who the person is, what the person does, and where the person is. For examples, recognition models can be exploited to identify people in pictures [15, 18], detection models can be used to detect human activities in videos [7, 22], and other inference models can be employed to infer individuals' locations in images [13, 30].

However, the performance of these visual-based models is greatly affected by the limited quality of visual data, and these models even will not be able to work when an object in pictures is occluded, when an activity occurs in the dark, or when an event happens in an area that is beyond the coverage of video cameras.

2.2 Audio-based Privacy Inference

Audio data can be used as a supplementary information source to achieve more stealthy privacy inference attacks own to its omnidirectional coverage and audio sensors' easy deployment in various environments [10, 11]. Therefore, a few research has begun to investigate the possibility of inferring privacy using audio data, which can be broadly classified into three mainstream applications. (i) Person identification can be accomplished by matching the newly captured timbre of a person from audio with the previously learned timbre of the same person [8, 9, 12, 23]. (ii) Vocal sounds produced by humans can also be recognized through audio data [1, 3, 4], which includes infants' and adults' screams, crying, coughing, clapping, whistling, sneezing, laughing, and the sound of footsteps. (iii) Indoor and outdoor environmental scenes where humans locate, such as homes, offices, and residential areas, can be detected by analyzing an audio stream as well [2, 6, 29, 31]. Although these existing works have demonstrated that it is possible to infer a single specific type of sensitive information about humans in audio, there is no one to design a scheme to directly speculate an integral event related to humans by analyzing polyphonic audio.

In this paper, three audio-based models are presented to realize the inference of human's integral event by processing polyphonic audio. The technical novelty of our models lies in two aspects. (i) The temporal correlation and the importance of different data points are leveraged in the sequential data feature learning. (ii) The co-occurrence dependency in multiple labels and the importance of these labels are exploited in the final event prediction.

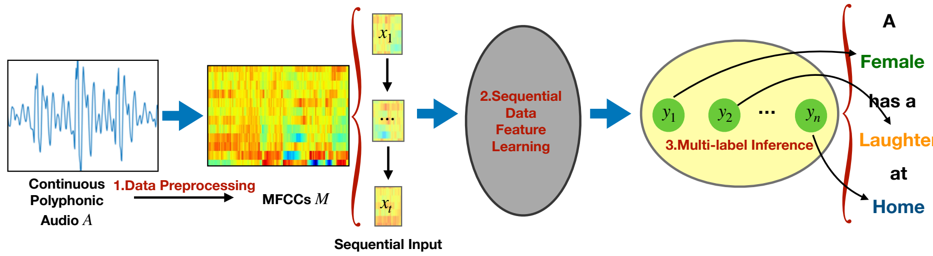


Fig. 1. Framework of Our Proposed Audio-based IntegraL Event Inference Model (ALTER)

3 METHODOLOGY

In this paper, we treat each element in an event as a label of one polyphonic audio. Accordingly, we aim to predict multiple labels of one polyphonic audio and then combine these labels related to the same event to infer the integral event. To this end, we propose an Audio-based integraL evenT infERence (ALTER) model as presented in Fig. 1. Generally speaking, ALTER is composed of three components, including (i) data preprocessing, (ii) sequential data feature learning, and (iii) multi-label inference. At the beginning, in data preprocessing, we convert the continuous polyphonic audio into Mel-Frequency Cepstrum Coefficients (MFCCs) [25]. Then, a sequential data feature learning scheme is used to capture the features of sequential input while considering the temporal correlation in the sequential data. Next, the multi-label inference stage leverages the extracted data features to predict multiple element labels. In the following, after introducing the design of three components of ALTER in Section 3.1, we present two upgraded models, ALTER-p and ALTER-pp, in Section 3.2 and Section 3.3, respectively.

3.1 ALTER

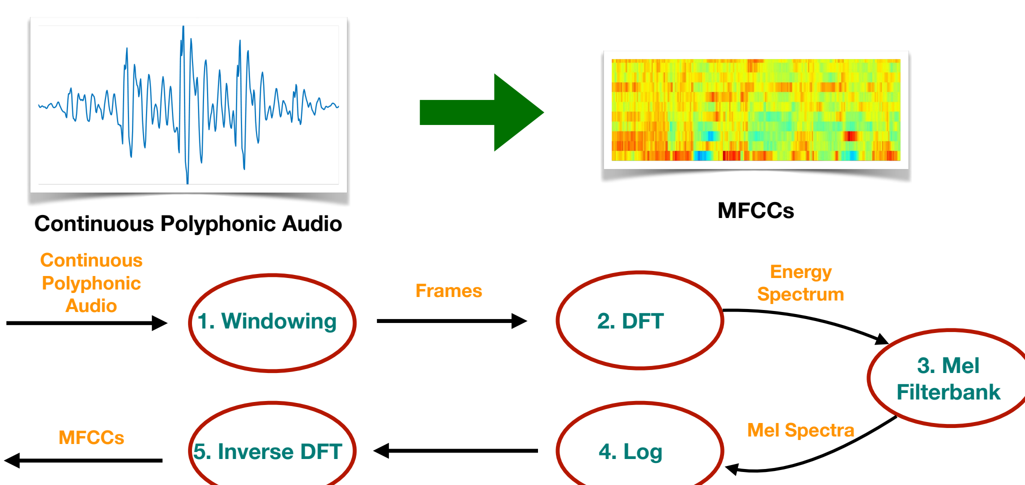


Fig. 2. Procedure of Calculating MFCCs

3.1.1 *Data Preprocessing.*

Since MFCCs have shown effectiveness in capturing the features of the acoustic signal in the speech recognition systems [19, 26, 33], we transform the polyphonic audio into MFCCs for our audio-based event inference task. In Fig. 2, we illustrate the procedure for calculating MFCCs of audio step by step: (i) window the original continuous polyphonic audio into a series of short frames; (ii) for each frame, calculate the energy spectrum using Discrete Cosine Transform (DCT) [32]; (iii) apply a mel filterbank [21], which is a series of bandpass filters with constant bandwidth and spacing on a mel frequency scale, to each frame's energy spectrum in order to get the multiple mel spectra; (iv) compute the logarithm of the mel spectra of each frame; and (v) convert these frames' logarithmic mel spectra back to the time domain via inverse DCT [32], which are MFCCs of the polyphonic audio. For presentation simplicity, we denote the calculation procedure of MFCCs as a function $F(\cdot)$ and use $F(\cdot)$ to transform the original continuous audio vector A_t into MFCCs matrix $M_{d\times t}$, *i.e.*,

$$M_{d\times t} = F(A_t), \tag{1}$$

where t is the dimension of the audio vector, and d is the number of filters in the filterbank.

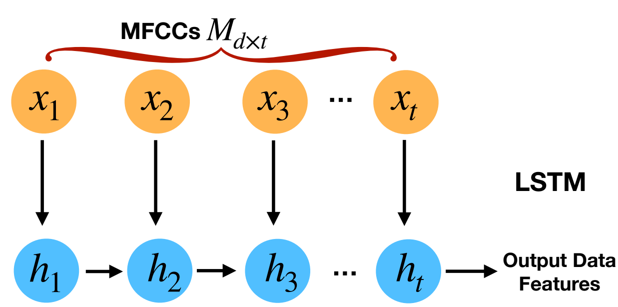


Fig. 3. LSTM-based Sequential Data Feature Learning

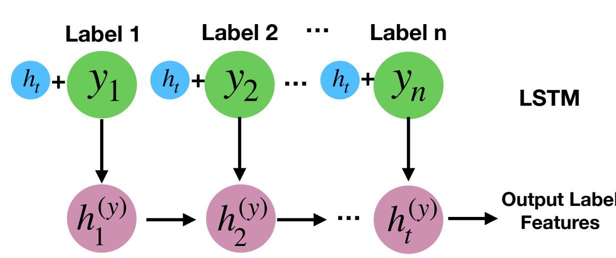


Fig. 4. LSTM-based Multi-label Feature Learning

3.1.2 *Sequential Data Feature Learning.* We treat the obtained MFCCs matrix as a sequence $M_{d\times t} = \{x_1, x_2, \cdots, x_t\}$, where each element is a d-dimensional vector. LSTM neural network [14] provides an extraordinary function to learn the features of sequential data with the consideration of temporal correlation in data. In light of this, we use the LSTM unit to extract the features from the sequence

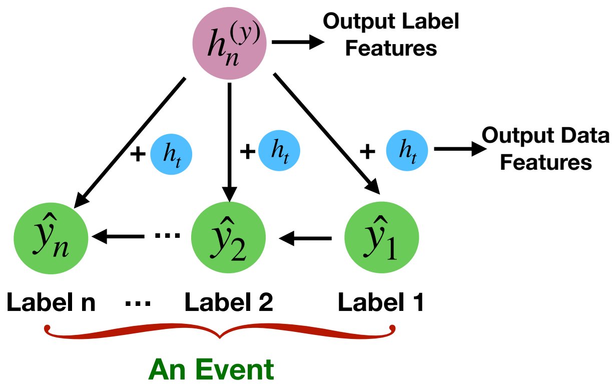


Fig. 5. Link-like Multi-label Inference

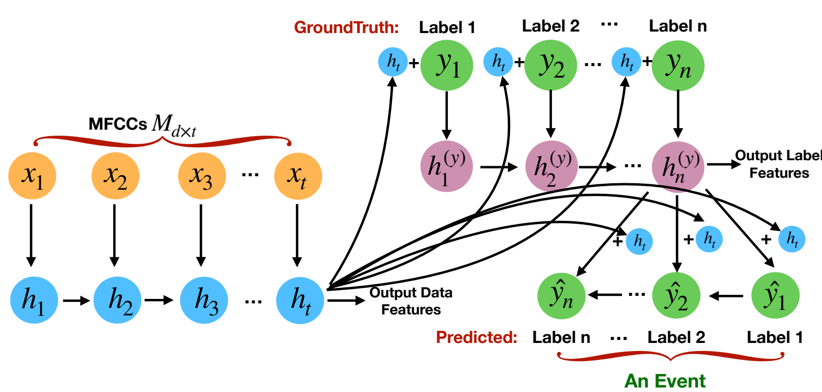


Fig. 6. Data Flow of ATLER Model

$M_{d\times t}$, which can be formulated as follows:

$$i_T = \sigma(W_i[h_{T-1}, x_T] + b_i), \tag{2}$$

$$f_T = \sigma(W_f[h_{T-1}, x_T] + b_f), \tag{3}$$

$$o_T = \sigma(W_o[h_{T-1}, x_T] + b_o), \tag{4}$$

$$\tilde{c}_T = \sigma(W_c[h_{T-1}, x_T] + b_c), \tag{5}$$

$$c_T = f_T c_{T-1} + i_T \tilde{c}_T, \tag{6}$$

$$h_T = o_T \cdot tanh(c_T), \tag{7}$$

where $x_T \in M_{d\times t}$; $T \in [2, t]$; i_T, f_T, and o_T are the input gate, forget gate, and output gate, respectively; $\sigma(\cdot)$ is the activation function; W_i, W_f, W_o, and W_c are the weights, and b_i, b_f, b_o, and b_c are the biases; $\tilde{c}_T$ is the immediate state, and c_T is the long-term state during sequential data feature learning process; $tanh(\cdot)$ is the hyberbolic tangent activation function; and x_T and h_T are T-th input and output information, respectively. The LSTM-based sequential data feature learning process is presented in Fig. 3, where we can get the final output features h_t from $M_{d\times t}$.

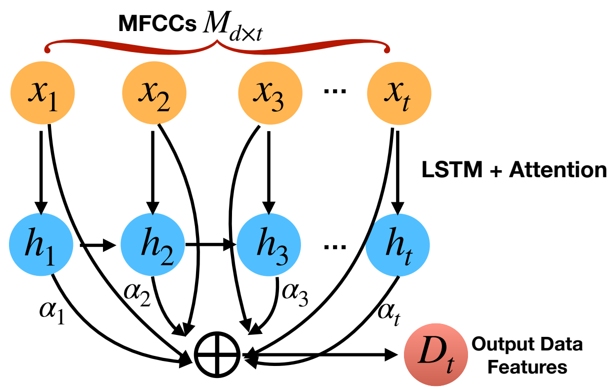


Fig. 7. LSTM-Attention-based Sequential Data Feature Learning

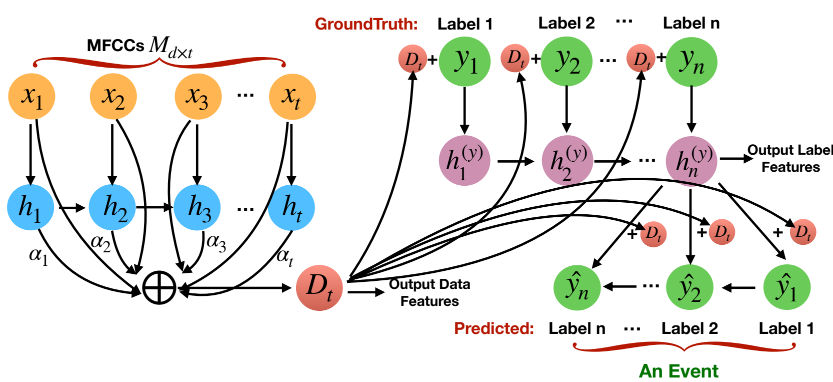


Fig. 8. Data Flow of ATLER-p Model

3.1.3 Multi-label Inference. It is known that an integral event can be described by several elements, such as who is a person, what is a person talking, and where is a person. In this paper, we assume that an integral event is composed of n elements, each of which can be taken as one label of continuous polyphonic audio. Thus, we can consider the event inference task as a multi-label inference task. Our multi-label inference process contains two phases, *i.e.*, multi-label feature learning and multi-label inference.

As a matter of fact, one event is usually composed of more than one concurrent element, including object, activity, environment, *etc.* For example, in an event that "a girl has a laughter at home", the coccurrent elements are gender (*i.e.*, female), activity (*i.e.*, laughter), and location (*i.e.*, home). That is, the elements in an event are co-occurrence dependent. Hence, for multi-label feature learning, we attempt to learn the features of multiple labels while considering the co-occurrence dependency among these element labels. We can treat these correlated labels as a label sequence and denote the label sequence as $Y = \{y_1, y_2, \cdots, y_n\}$, where y_i is the i-th element label in the event. In Fig. 4, we exploit LSTM neural network to extract the multi-label features with incorporating label correlation. Furthermore, taking into account that the data features mainly affect the labels' prediction, the output data features h_t are also used in our LSTM-based multi-label feature learning, which can be formulated below:

$$i_N^{(y)} = \sigma(W_i^{(y)}[h_{N-1}^{(y)}, y_N] + A_i h_t + b_i^{(y)}), \tag{8}$$

$$f_N^{(y)} = \sigma(W_f^{(y)}[h_{N-1}^{(y)}, y_N] + A_f h_t + b_f^{(y)}), \tag{9}$$

$$o_N^{(y)} = \sigma(W_o^{(y)}[h_{N-1}^{(y)}, y_N] + A_o h_t + b_o^{(y)}), \tag{10}$$

$$\tilde{c}_N^{(y)} = \sigma(W_c^{(y)}[h_{N-1}^{(y)}, y_N] + A_c h_t + b_c^{(y)}), \tag{11}$$

$$c_N^{(y)} = f_N^{(y)} c_{N-1}^{(y)} + i_N^{(y)} \tilde{c}_N^{(y)}, \tag{12}$$

$$h_N^{(y)} = o_N^{(y)} \cdot tanh(c_N^{(y)}), \tag{13}$$

where $y_N \in Y$; $N \in [2, n]$; $i_N^{(y)}$ $f_N^{(y)}$, and $i_N^{(y)}$ are the input gate, forget gate, and output gate for label feature learning, respectively; $W_i^{(y)}$, $W_f^{(y)}$, $W_o^{(y)}$, and $W_c^{(y)}$ are the weights, and $b_i^{(y)}$, $b_f^{(y)}$, $b_o^{(y)}$, and $b_c^{(y)}$ are the biases in the label feature learning process; $\tilde{c}_N^{(y)}$ is the immediate state, and $c_N^{(y)}$ is the long-term state during the label feature learning; y_N and $h_N^{(y)}$ are N-th input and output label information, respectively; and A_i, A_f, A_o, and A_c are the weights of data features in the LSTM-based label feature learning architecture. Consequently, we can obtain the final label features $h_n^{(y)}$ for further inference.

Moreover, as presented in Fig. 5, we propose a link-like multi-label inference, during which we consider the fact that the current predicted label $\hat{y_N}$ can be influenced by the previous one predicted label $\hat{y}_{N-1}$, the output data features h_t, and output label features $h_n^{(y)}$. So, we design the final layer using $softmax(\cdot)$ function shown in Eq. (14).

$$\hat{y}_N = softmax(U_s \sigma(W_s[h_n^{(y)}, h_t, \hat{y}_{N-1}]) + b_s), \tag{14}$$

where U_s, W_s, b_s are the parameters of $softmax(\cdot)$ to be learned.

At the end, we present the data flow of our proposed ATLER model in Fig. 6 by combining the aforementioned three components. The ALTER model is trained by minimizing the summation of the cross entropy between the predicted label $\hat{y}_n$ and the corresponding ground-truth label y_n.

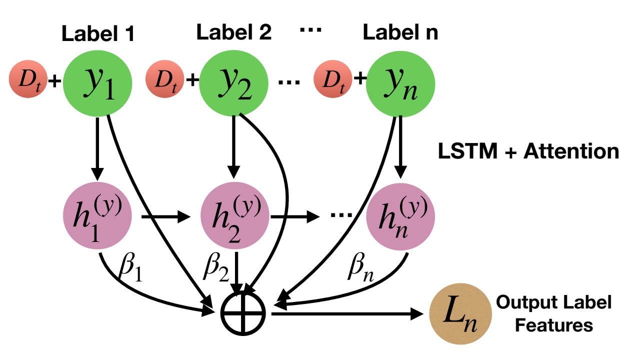


Fig. 9. LSTM-Attention-based Multi-label Feature Learning

3.2 ALTER-p

In ALTER, we use LSTM to extract the data features to get the output h_t, which, however, compresses too much original data information. In order to make full use of all data information and the

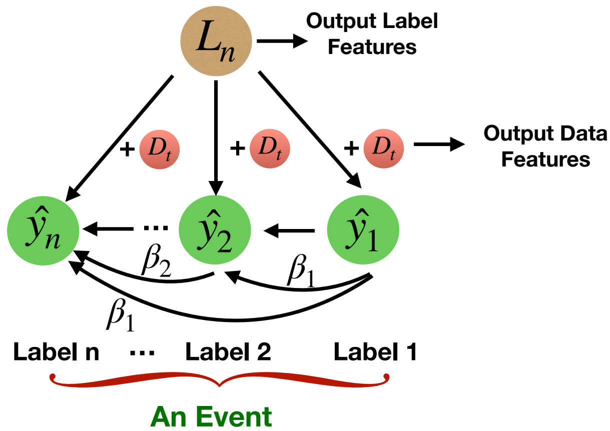


Fig. 10. Graph-like Multi-label Inference

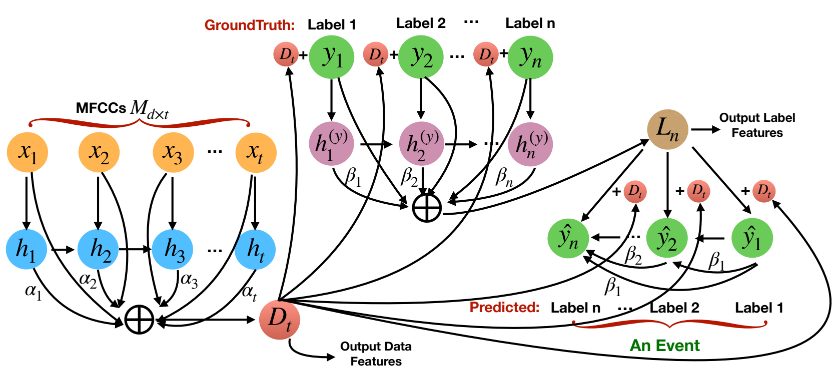


Fig. 11. Data Flow of ATLER-pp Model

importance of data points at the same time, we update our sequential data feature learning by using LSTM and attention mechanisms simultaneously. In Fig. 7, we show the LSTM-Attention-based sequential data feature learning scheme. First of all, we can compute the unnormalized relevance score e_T of the data x_T by using the following attention function:

$$e_T = Q_e \cdot tanh(W_e h_T + U_e x_T + z_e), \tag{15}$$

where $T \in [1, t]$, and Q_e, W_e, U_e, and z_e are the parameters in the attention function. Then, we can calculate the corresponding attention weight α_T in Eq. (16) via normalizing the relevance scores.

$$\alpha_T = exp(e_T) / \sum_{T=1}^{t} exp(e_T). \tag{16}$$

Based on these attention weights, we define the new output data features as:

$$D_t = \sum_{T=1}^{t} \alpha_T x_T. \tag{17}$$

Accordingly, the final prediction function in Eq. (14) should be updated as:

$$\hat{y}_N = softmax(U_s\sigma(W_s[h_n^{(y)}, D_t, \hat{y}_{N-1}]) + b_s). \tag{18}$$

Finally, we replace the original LSTM-based one in ALTER with the LSTM-Attention-based sequential data learning to obtain our ALTER-p model, the data flow of which is shown in Fig. 8. ALTER-p is be trained in the same way as the ALTER model.

3.3 ALTER-pp

Similarly, we expect to obtain the label features by using all label information while considering the importance of multiple labels. For this purpose, the LSTM-Attention architecture shown in Fig. 9 is applied to update our original LSTM-based multi-label feature learning component. In this architecture, we first calculate the unnormalized relevance score $e_N^{(y)}$ of the label y_N, *i.e.*,

$$e_N^{(y)} = Q_e^{(y)} \cdot tanh(W_e^{(y)} h_N^{(y)} + U_e^{(y)} y_N + z_e^{(y)}), \tag{19}$$

where $N \in [1, n]$, and $Q_e^{(y)}$, $W_e^{(y)}$, $U_e^{(y)}$, and $z_e^{(y)}$ are the parameters of the attention function in label feature learning. Then, the attention weight of the N-th label β_N can be computed as:

$$\beta_N = exp(e_N^{(y)}) / \sum_{N=1}^{n} exp(e_N^{(y)}). \tag{20}$$

Consequently, we define the new label features to be:

$$L_n = \sum_{N=1}^{n} \beta_N y_N. \tag{21}$$

Moreover, inspired by the attention-based learning process, we propose a new graph-like multi-label inference presented in Fig. 10, where the prediction result of current label $\hat{y_N}$ is affected by all previously predicted labels $\{\hat{y}_1, \cdots, \hat{y}_{N-1}\}$. Thus, by using the newly learned label features L_n and the graph-like multi-label inference idea, we can further improve the prediction function in Eq. (22).

$$\hat{y}_N = softmax(U_s\sigma(W_s[L_n, D_t, \sum_{j=1}^{N-1} \beta_j \hat{y}_j]) + b_s). \tag{22}$$

After all, ALTER-pp is constructed by employing LSTM-Attention-based sequential data feature learning, LSTM-Attention-based multi-label feature learning, and graph-like multi-label inference, the data flow of which is demonstrated in Fig. 11. We will also train ALTER-pp using the same way of training ALTER.

Table 1. Gender Prediction Results (Ours v.s. Baseline 1)

Model	Data Learning	Label Learning	Acc	Pre	Rec	F1	Auc
Baseline 1	/	/	0.834	0.902	0.698	0.832	0.916
ALTER	**LSTM**	**LSTM**	0.844 (↑ 1.20%)	0.911 (↑ 1.00%)	0.699 (↑ 0.14%)	0.842 (↑ 1.20%)	0.918 (↑ 0.22%)
ALTER-p	**LSTM + Attention**	**LSTM**	0.845 (↑ 1.32%)	0.916 (↑ 1.55%)	0.707 (↑ 1.29%)	0.843 (↑ 1.32%)	0.922 (↑ 0.66%)
ALTER-pp	**LSTM + Attention**	**LSTM + Attention**	0.846 (↑ 1.44%)	0.919 (↑ 1.88%)	0.718 (↑ 2.87%)	0.844 (↑ 1.44%)	0.926 (↑ 1.09%)

Table 2. Vocal Sound Prediction Results (Ours v.s. Baseline 2)

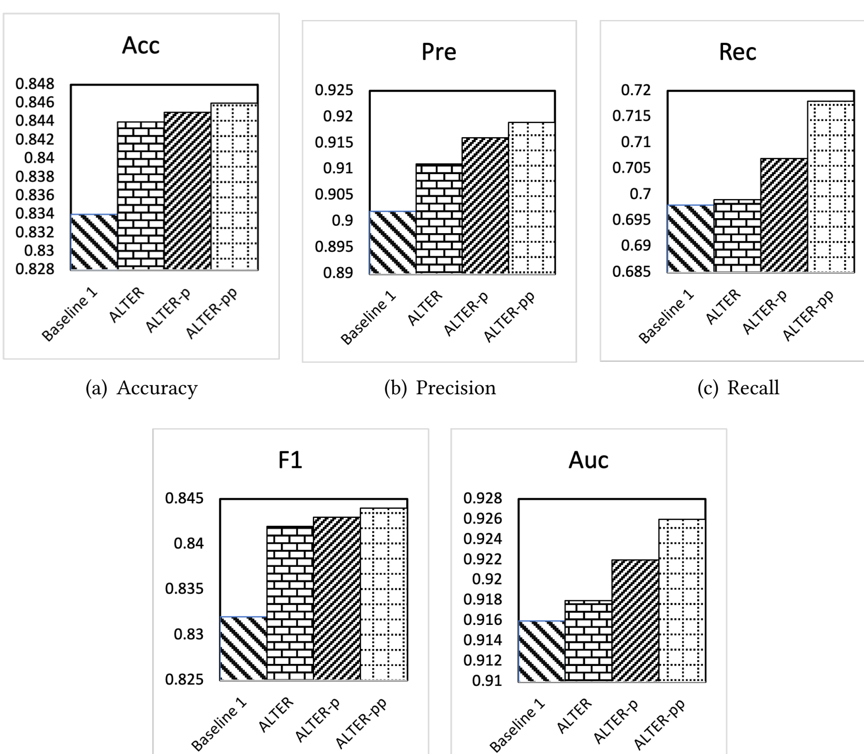
Model	Data Learning	Label Learning	Acc	Pre	Rec	F1	Auc
Baseline 2	/	/	0.900	0.925	0.800	0.901	0.975
ALTER	**LSTM**	**LSTM**	0.908 (↑ 0.89%)	0.944 (↑ 2.05%)	0.833 (↑ 4.13%)	0.908 (↑ 0.78%)	0.980 (↑ 0.51%)
ALTER-p	**LSTM + Attention**	**LSTM**	0.914 (↑ 1.56%)	0.949 (↑ 2.59%)	0.836 (↑ 4.50%)	0.914 (↑ 1.44%)	0.986 (↑ 1.13%)
ALTER-pp	**LSTM + Attention**	**LSTM + Attention**	0.918 (↑ 2.00%)	0.954 (↑ 3.14%)	0.898 (↑ 12.25%)	0.918 (↑ 1.89%)	0.993 (↑ 1.85%)

Table 3. Environment Prediction Results (Ours v.s. Baseline 3)

Model	Data Learning	Label Learning	Acc	Pre	Rec	F1	Auc
Baseline 3	/	/	0.987	0.964	0.503	0.979	0.976
ALTER	**LSTM**	**LSTM**	0.989 (↑ 0.20%)	0.975 (↑ 1.14%)	0.511 (↑ 1.59%)	0.984 (↑ 0.51%)	0.985 (↑ 0.92%)
ALTER-p	**LSTM + Attention**	**LSTM**	0.997 (↑ 1.01%)	0.985 (↑ 2.18%)	0.534 (↑ 6.16%)	0.994 (↑ 1.53%)	0.993 (↑ 1.74%)
ALTER-pp	**LSTM + Attention**	**LSTM + Attention**	0.998 (↑ 1.11%)	0.988 (↑ 2.49%)	0.567 (↑ 12.72%)	0.997 (↑ 1.84%)	0.995 (↑ 1.95%)

Table 4. Event Prediction Results (Ours v.s. Baseline)

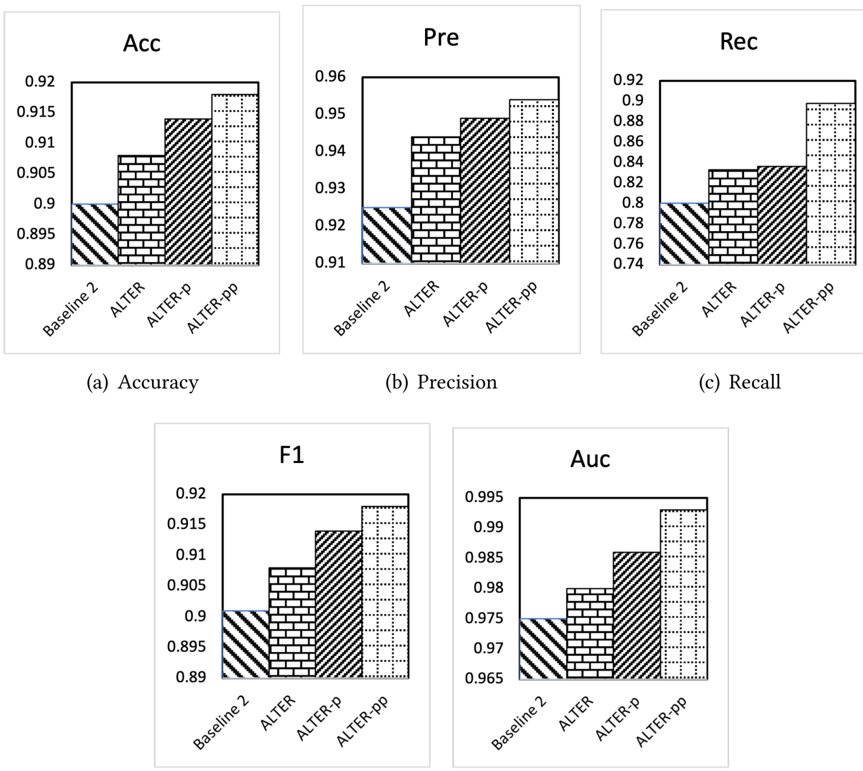
Model	Data Learning	Label Learning	Acc	Pre	Rec	F1	Auc
Baseline	/	/	0.718	0.521	0.828	0.704	0.945
ALTER	**LSTM**	**LSTM**	0.734	0.534	0.890	0.729	0.966
ALTER-p	**LSTM + Attention**	**LSTM**	0.778 (↑ 5.99%)	0.587 (↑ 9.93%)	0.896 (↑ 0.67%)	0.776 (↑ 6.45%)	0.970 (↑ 0.41%)
ALTER-pp	**LSTM + Attention**	**LSTM + Attention**	0.781 (↑ 6.40%)	0.675 (↑ 26.4%)	0.897 (↑ 0.79%)	0.779 (↑ 6.96%)	0.975 (↑ 0.93%)



(a) Accuracy (b) Precision (c) Recall

(d) F1 Score (e) Auc

Fig. 12. Gender Prediction Results (Ours v.s. Baseline 1)



(a) Accuracy (b) Precision (c) Recall

(d) F1 Score (e) Auc

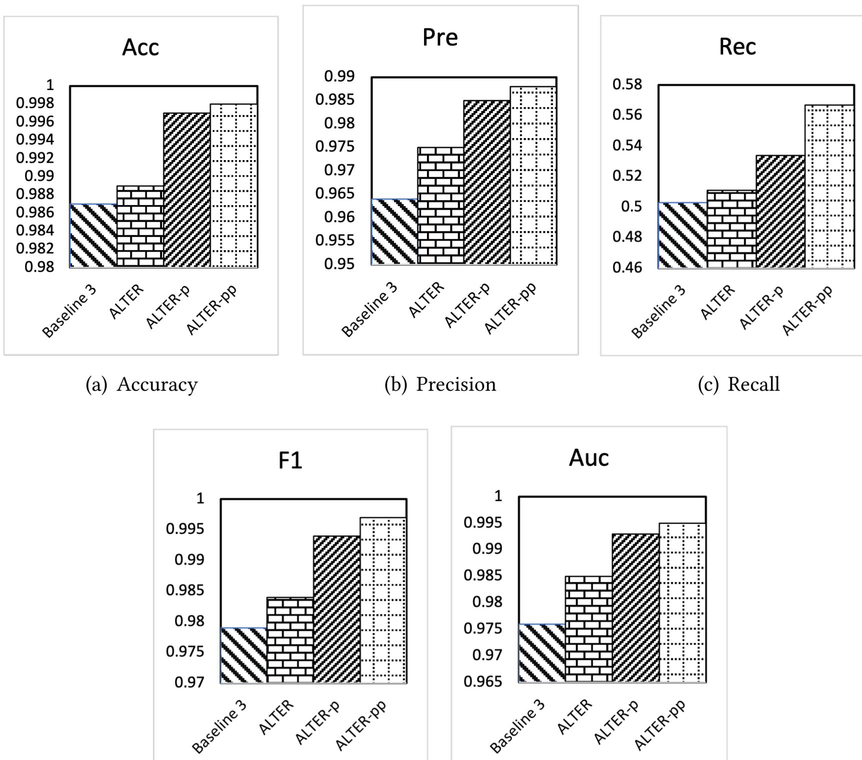
Fig. 13. Vocal Sound Prediction Results (Ours v.s. Baseline 2)

3.4 Model Comparison

We design ALTER model to infer an audio-based integral event by leveraging the temporal correlation in audio and the co-occurrence dependency among multiple element labels. However, in ATLER, the LTSM-based sequential data feature learning, which compresses the audio data into the output data features, may lead to data information loss when processing relatively longer audio. To reduce such information loss, ALTER-p is proposed by making full use of audio information and the importance of all data points, which is more helpful to analyze an audio with a relatively longer time period. Nonetheless, in the link-like multi-label inference of ALTER-p, we only consider a short-term co-occurrence dependency among labels, which may be limited in predicting a complicated event with relatively more elements. While, in order to effectively predict a sophisticated event with diverse elements, ALTER-pp is further presented by taking advantage of the long-term co-occurrence dependency among labels (*i.e.*, the graph-like multi-label inference).

4 EXPERIMENTS

In this section, we first introduce the experiment settings and then conduct comprehensive experiments to evaluate the effectiveness of our proposed ALTER, ALTER-p, and ALTER-pp models on a real-world dataset. Besides, more extensive experiments are done to compare our proposed models with the state-of-the-art.



(a) Accuracy (b) Precision (c) Recall

(d) F1 Score (e) Auc

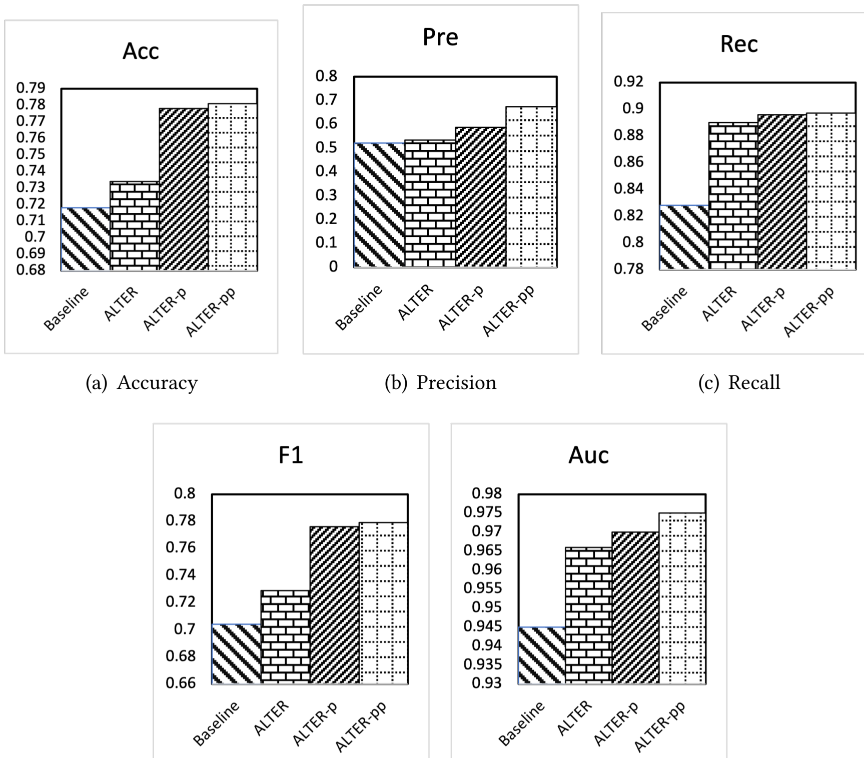
Fig. 14. Environment Prediction Results (Ours v.s. Baseline 3)

4.1 Experiment Settings

The datasets, baselines, performance metrics, network architectures, and parameter settings are described below.

4.1.1 Dataset. We adopt two public datasets, including **VocalSound** [17] and **TUT Acoustic Scenes 2016** [24]. VocalSound is a dataset consisting of males' and females' recordings of "laughter, sigh, cough, throat clearing, sneeze, and sniff". TUT Acoustic Scenes 2016 includes recordings from various acoustic environments, such as homes, offices, and residential areas. Since we aim to test the performance of our audio-based integral event inference models in the experiments, we synthesize these two datasets to obtain a polyphonic audio dataset, which contains human gender information, human vocal sound information, and environmental information. In this synthetic dataset, for instance, one polyphonic audio records an event that "a female has a laughter at home", and the corresponding labels of this audio record are "female", "laughter", and "home".

4.1.2 Baselines. Although no work has been proposed to predict an integral event based on audio so far, there are some related works to infer one element in an event. The one-element event inference can be treated as a special case in our models. Thus, we choose the following baselines to conduct comparison experiments so as to further illustrate the superiority of our models in this special case. (1) An EfficientNet-based model proposed in [17] is a state-of-the-art model for the



(a) Accuracy (b) Precision (c) Recall

(d) F1 Score (e) Auc

Fig. 15. Event Prediction Results (Ours v.s. Baseline)

gender prediction on VocalSound dataset. (2) In [17], another state-of-the-art EfficientNet-based approach is presented to make human vocal sound inference on VocalSound dataset. (3) A GMM-based model in [24] is the state-of-the-art to achieve environment recognition on TUT Acoustic Scenes 2016 dataset.

4.1.3 Performance Metrics. Since the audio-based integral event inference can be considered as a multi-label classification task, we use five typical metrics for classification tasks as the performance measurements, including accuracy (Acc), precision (Pre), recall (Rec), F1 score (F1), and area under the receiver operating characteristic curve (Auc). A higher value of Acc indicates a more precise prediction outcome, and the same principle applies to Pre, Rec, F1, and Auc.

4.1.4 Network Architectures. In ALTER model, we use two LSTM layers for sequential data feature learning and another two LSTM layers in the multi-label inference phase. For ALTER-p, we maintain the design of multi-label inference in ALTER and update the sequential data feature learning in ALTER by applying two LSTM layers and an attention layer concurrently. For ALTER-pp, we follow the sequential data feature learning architecture in ALTER-p while achieving multi-label inference via two LSTM layers plus an attention layer.

4.1.5 Parameter Settings. In data preprocessing, we window each audio sample into short frames every 10*ms* and use a filterbank with 128 filters to convert the audio vector into the MFCCs matrix.

We train the neural networks in our proposed ALTER, ALTER-p, and ALTER-pp models using an Adam optimizer for 80 epochs with an initial learning rate at $1e-4$ and a batch size of 100.

4.2 Comparison between Ours and Baselines

In order to verify the effectiveness of ALTER, ALTER-p, and ALTER-pp models on the one-element event inference, we compare the performance of our proposed models with three state-of-the-art baselines. Firstly, we show the gender recognition results of our models and baseline 1 in Table 1 and Fig. 12, where it can be seen that our proposed models' performance is comparable and even better than baseline 1. Secondly, the vocal sound prediction results of our models and the baseline 2 are presented in Table 2 and Fig. 13. From these results, we can find out that the proposed models outperform baseline 2 with regard to human vocal sound inference. Thirdly, by comparing the results of Acc, Pre, Rec, F1, and Auc in Table 3 and Fig. 14, we can notice that our models are superior to baseline 3 in terms of environment prediction. To sum up, our models have superiority over the previous state-of-the-art approaches in terms of one specific element inference since the temporal correlation and the importance of different data points are leveraged in our proposed sequential data feature learning. The names of the baselines and their corresponding models are shown in Table 5.

Table 5. Baseline Models

Baseline 1	EfficientNet-based Model [17]
Baseline 2	EfficientNet-based Model [17]
Baseline 3	GMM-based Model [24]

4.3 Evaluation on Our Models

Since the problem of integral event inference has not been address by existing works, we combine baseline 1, baseline 2, and baseline 3 to obtain a event inference model, which is used as a baseline to investigate the effectiveness of our proposed models. To be specific, after training our models and the baseline model, we use the trained models to test the polyphonic audios in the testing dataset to predict the multiple element labels. Then, the predicted element labels and the corresponding ground-truth ones are used to calculate the event prediction performance to measure the event inference effectiveness, for which we present the values of Acc, Pre, Rec, F1, and Auc in Table 4 and Fig. 15. The results demonstrate that ALTER model outperforms the baseline in terms of integral event inference on polyphonic audio thanks to the incorporation of the temporal correlation in audio and the short-term co-occurrence dependency among multiple labels simultaneously. Besides, by comparing ALTER-p with ALTER, we can see that the values of all performance metrics are increased. Significantly, Acc and F1 are increased by about 6.00%, and Pre is increased by about 10.00%. The comparison indicates that ALTER-p can enhance the performance of the event prediction due to the full utilization of data information and the importance of all data points. In addition, compared with ALTER-p, ALTER-pp can obtain more improvements in the event prediction performance thanks to the consideration of the long-term co-occurrence dependency among labels.

5 DISCUSSION AND FUTURE WORK

In this section, we discuss two limitations of this work and present our future research directions.

(i) Although the experimental results have shown that ALTER-p can improve the performance of event prediction by considering the whole data information and the importance of all the data points, the performance improvement is not too much since the audio samples in our synthetic dataset are short. Therefore, in the future, it is desirable for us to highlight the advantage of ALTER-p by collecting longer real-world audios via extensive experiments.

(ii) Similarly, due to the limitation of data source, we use our ALTER-pp model to predict the three-element event. As a result, the graph-like multi-label inference in ALTER-pp cannot bring too much performance improvement. We will conduct more comprehensive experiments after collecting polyphonic audios of human events with more diverse elements so as to better evaluate the benefit of considering the long-term co-occurrence dependency among labels.

6 CONCLUSION

This paper is the first work to investigate an audio-based integral event inference. Firstly, we propose an ALTER model to effectively achieve event inference by leveraging the temporal correlation in audio and the short-term co-occurrence dependency among multiple labels. Moreover, ALTER-p is designed by fully exploiting data information and the importance of all data points so as to enhance event prediction performance. Furthermore, ALTER-pp is proposed by further considering the long-term co-occurrence dependency among multiple labels for event inference performance improvement. Finally, via comprehensive real-data experiments, we demonstrate the effectiveness of our proposed models on the integral event inference and their advantages over the state-of-the-art methods.

ACKNOWLEDGMENTS

This work was partly supported by the National Science Foundation of U.S. (2146497, 2416872, 2315596, 2244219, 2416871, 2244220, 2343619).

REFERENCES

[1] Sharath Adavanne and Tuomas Virtanen. 2017. A report on sound event detection with different binaural features. *arXiv preprint arXiv:1710.02997* (2017).

[2] Dharmesh M Agrawal, Hardik B Sailor, Meet H Soni, and Hemant A Patil. 2017. Novel TEO-based Gammatone features for environmental sound classification. In *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 1809–1813.

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems* 29 (2016).

[4] Rohan Badlani, Ankit Shah, Benjamin Elizalde, Anurag Kumar, and Bhiksha Raj. 2018. Framework for evaluation of sound event detection in web videos. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3096–3100.

[5] Sergio Benini, Khalil Khan, Riccardo Leonardi, Massimo Mauro, and Pierangelo Migliorati. 2019. Face analysis through semantic face segmentation. *Signal Processing: Image Communication* 74 (2019), 21–31.

[6] Venkatesh Boddapati, Andrej Petef, Jim Rasmusson, and Lars Lundberg. 2017. Classifying environmental sounds using image recognition networks. *Procedia computer science* 112 (2017), 2048–2056.

[7] Paulo Vinicius Koerich Borges, Nicola Conci, and Andrea Cavallaro. 2013. Video-based human behavior understanding: A survey. *IEEE transactions on circuits and systems for video technology* 23, 11 (2013), 1993–2008.

[8] Alessio Brutti and Andrea Cavallaro. 2017. Unsupervised cross-modal deep-model adaptation for audio-visual re-identification with wearable cameras. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. IEEE, 438–445.

[9] Andrea Cavallaro and Alessio Brutti. 2019. Audio-visual learning for body-worn cameras. In *Multimodal Behavior Analysis in the Wild*. Elsevier, 103–119.

[10] Jyotismita Chaki. 2021. Pattern analysis based acoustic signal processing: a survey of the state-of-art. *International Journal of Speech Technology* 24, 4 (2021), 913–955.

[11] S Chandrakala and SL Jayalakshmi. 2019. Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. *ACM Computing Surveys (CSUR)* 52, 3 (2019), 1–34.

[12] François-Xavier Decroix, Julien Pinquier, Isabelle Ferrané, and Frédéric Lerasle. 2016. Online audiovisual signature training for person re-identification. In *Proceedings of the 10th International Conference on Distributed Smart Camera*.

ACM, 62–68.

[13] Meriem Djouadi and Mohamed-Khireddine Kholladi. 2022. Improving Street View Image Classification Using Pre-trained CNN Model Extracted Features. *Periodica Polytechnica Electrical Engineering and Computer Science* 66, 4 (2022), 370–379.

[14] Furkan Elmaz, Reinout Eyckerman, Wim Casteels, Steven Latré, and Peter Hellinckx. 2021. CNN-LSTM architecture for predictive indoor temperature modeling. *Building and Environment* 206 (2021), 108327.

[15] Markus Enzweiler and Dariu M Gavrila. 2008. Monocular pedestrian detection: Survey and experiments. *IEEE transactions on pattern analysis and machine intelligence* 31, 12 (2008), 2179–2195.

[16] Sakher Ghanem, Ashiq Imran, and Vassilis Athitsos. 2019. Analysis of hand segmentation on challenging hand over face scenario. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 236–242.

[17] Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A Dataset for Improving Human Vocal Sounds Recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 151–155.

[18] Sven Gotovac, Danijel Zelenika, Željko Marušić, and Dunja Božić-Štulić. 2020. Visual-based person detection for search-and-rescue with uas: Humans vs. machine learning algorithm. *Remote Sensing* 12, 20 (2020), 3295.

[19] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, and Kong-Pang Pun. 2006. An efficient MFCC extraction method in speech recognition. In *2006 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 4–pp.

[20] Ruochen Jiang, Changbo Qu, Jiannan Wang, Chi Wang, and Yudian Zheng. 2020. Towards extracting highlights from recorded live videos: An implicit crowdsourcing approach. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1810–1813.

[21] Sunil Kumar Kopparapu and M Laxminarayana. 2010. Choice of Mel filter bank in computing MFCC of a resampled speech. In *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*. IEEE, 121–124.

[22] Honghai Liu, Zhaojie Ju, Xiaofei Ji, Chee Seng Chan, and Mehdi Khoury. 2017. Study of human action recognition based on improved spatio-temporal features. In *Human Motion Sensing and Recognition*. Springer, 233–250.

[23] Mirko Marras, Pedro A Marín-Reyes, Javier Lorenzo-Navarro, Modesto Castrillón-Santana, and Gianni Fenu. 2019. Deep multi-biometric fusion for audio-visual user re-identification and verification. In *International Conference on Pattern Recognition Applications and Methods*. Springer, 136–157.

[24] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 1128–1132.

[25] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083* (2010).

[26] Seiichi Nakagawa, Longbiao Wang, and Shinji Ohtsuka. 2011. Speaker identification and verification by combining MFCC and phase information. *IEEE transactions on audio, speech, and language processing* 20, 4 (2011), 1085–1095.

[27] Yan Pin Ong, Chun Meng Tan, and Carmen Jia Yi Siau. 2019. Systems and methods for processing a video stream during live video sharing. US Patent 10,412,318.

[28] Takashi Onohara, Roka Ueda, Keishi Daini, Taichi Yoshio, Yuji Kawabe, Seizi Iwayagano, HIGO Takuma, and Eri Sakai. 2019. Information-sharing device, method, and terminal device for sharing application information. US Patent 10,282,316.

[29] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 1015–1018.

[30] Xueming Qian, Huan Wang, Yisi Zhao, Xingsong Hou, Richang Hong, Meng Wang, and Yuan Yan Tang. 2016. Image location inference by multisaliency enhancement. *IEEE Transactions on Multimedia* 19, 4 (2016), 813–821.

[31] Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters* 24, 3 (2017), 279–283.

[32] Julius Orion Smith. 2008. *Mathematics of the discrete Fourier transform (DFT): with audio applications*. Julius Smith.

[33] Vibha Tiwari. 2010. MFCC and its applications in speaker recognition. *International journal on emerging technologies* 1, 1 (2010), 19–22.

[34] Sagar Kumar Verma. 2020. Method and system for sharing an output device between multimedia devices to transmit and receive data. US Patent 10,581,933.