Salient sound extraction using deep neural networks predicting complex masks

1st Tomasz Grzywalski, 2nd Dick Botteldooren
Department of Information Technology
Ghent University
Ghent, Belgium
0000-0002-9388-0494 0000-0002-7756-7238

3rd Yanjue Song, 4th Nilesh Madhu
IDLab, Department of Electronics and Information Systems
Ghent University - imec
Ghent, Belgium

***Abstract*—This work addresses the problem of extracting sounds that are unexpected in an audio stream and stand out because of their spectrotemporal characteristics. In human auditory scene analysis, such sounds are referred to as (sensory) salient. Previous research initiatives are mostly limited to the detection of presence of salient sounds and identification of their temporal localization within the signal. Other approaches aim at developing classifiers that detect fixed, predetermined categories of salient sounds. In contrast, this work aims at developing a solution capable of suppressing all background (non-salient) sounds from an audio stream, preserving, to the best extent possible, the salient sounds without any distortion. An additional assumption is that the algorithm should not be limited to any particular category of salient sound events. This challenging task is realized in two steps, both being novel contributions of this work. In the first step, a large-scale dataset of clean background samples and clean salient sound samples is created by automatically processing publicly available resource of field recordings. In the second step, a deep neural network (U-Net) trained to predict complex ideal ratio mask, a method typically used for speech enhancement, is adopted and evaluated in the context of salient sound extraction. The results of conducted experiments indicate potential high efficacy of the proposed solution and indicate directions for future research.**

***Index Terms*—salient sound extraction, salient event detection, deep neural networks**

I. Introduction

This work deals with the problem of extraction of salient sound events from any audio stream. In this context, salient sounds are defined as sounds that stand out from the ongoing stream because of their unexpected spectrotemporal features. These could be sudden events in a more continuous stream but also events of a different nature in a stream of events. In human auditory scene analysis, such detections are sometimes referred to as sensory salient sounds to distinguish them from cognitive or semantic salience where a recognized sound stands out because it does not fit or is not expected in a particular sound environment [1], [2]. The broad range of environmental sounds requires considering different time scales from seconds to hours, but although the proposed methods would also be applicable on very slow or very fast events, we focus this manuscript on what a human observer would commonly call an event or a transition in the sound stream.

Aiming at generic saliency detection and extraction is a difficult task as it is impossible to include a priori assumptions regarding the nature of the acoustical environment in which the system will be working and therefore also about what could be considered salient in this context. Instead, we want the model to adapt to any situation and be able to differentiate between sounds that belong to the acoustical background and therefore are considered normal and sound events that stand out from this norm. Such a system, when developed, could have a wide range of applications but would be most useful in serving as a pre-processing step in passive acoustic monitoring tasks, such as audio surveillance of public spaces or industrial sites, monitoring of wildlife, livestock, or machine fault detection. With such a system, all downstream processing could be done only when the salient sounds are detected and de-noised, which should render the whole pipeline time- and resource-effective. However, to achieve this goal, the salient sound extraction module must itself be computationally inexpensive and accurate in the sense that it must not distort the salient sounds so that they can be identified and acted upon in the downstream modules.

Historically, many attempts have been made to detect rare/anomalous/salient sounds in an audio signals [3], [4]. However, in most previous works, types of acoustic scenes and/or salient sound types were very limited [5]–[7]. The proposed system differs from these approaches in two key areas: (1) we want our model to be able to detect all salient sound, not only a predefined set of sound types and (2) we are aiming at extracting the salient sound events from the input signal so that they can be further processed depending on the application.

Currently, the development of such a system is challenging due to the lack of datasets of salient sounds that would exhibit sufficient scale and diversity to achieve the outlined goal. To address the data scarcity problem, which is otherwise difficult to overcome [8], we have curated a large-scale dataset of salient-free background recordings and an equally large set of clean salient sounds. The first part of this work describes in detail the process of constructing the dataset.

In the second part, we adopt a state-of-the-art solution for monaural speech enhancement [9]–[11] to the problem of salient sound event extraction. This solution involves the usage of an U-shaped deep neural network (U-Net, [12]) trained on sound mixtures to predict a complex ideal ratio mask (cIRM,

[13]), which, when applied to the input mixture, performs the background noise suppression. To the best of our knowledge, such an application has not been attempted before according to the literature. Extensive evaluation using an in-house test set is conducted to test the effectiveness of the proposed solution both in the context of the detection of the presence of salient sound events in the audio as well as its ability to suppress background noise with minimal distortion, thus allowing for the automatic identification of the salient sound event type by an external system.

II. Dataset of salient sounds

To train a general-purpose salient sound extraction neural network, a high-quality, diverse dataset of sounds is needed. The dataset should consist of two sets of recordings: (1) background recordings, which are known not to contain any salient sound events, and (2) recordings of isolated sounds that become salient when mixed with a random background.

The dataset was constructed by first downloading a large set of unlabeled field recordings from an online database. Next, recordings with no salient sounds were identified using a sound event detection neural network with a short effective receptive field. Finally, isolated salient sounds were extracted from recordings that contained them. The following sections describe the process of development of this dataset, which is summarized in Fig 1.

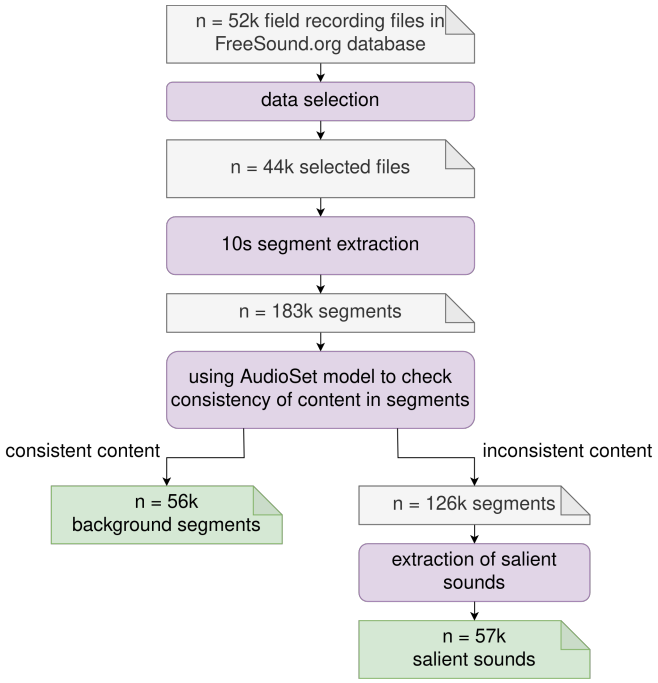


Fig. 1. Summary of the salient sound dataset creation process.

A. Field recordings

In the first step of dataset creation, a large and diverse set of unlabeled field recordings was gathered from freesound.org [14]. The process was conducted in July 2022 when the online resource contained 51599 recordings tagged with 'field recording' tag. From this set, all recordings longer than 10 seconds were downloaded. Very long recordings were skipped to allow the downloading to finish in a reasonable time (approximately one week). Eventually, 43986 recordings (85%) were downloaded.

Next, all downloaded recordings were resampled to 8kHz and up to 5 non-overlapping, 10-second segments were extracted from each recording. This yielded a total of 182646 segments of a total duration of over 507 hours, an average of 4.2 segments per downloaded recording.

To gain insight into the content of these segments, they were processed with a popular sound classification model, PANN [15], which recognizes 527 classes of sounds. The PANN classified the segments into 405 classes, from which 119 were represented by more than 100 segments. The most prevalent classes were: Speech (32777 segments, 17.9%), Vehicle (22210, 12.2%), Bird (17305, 9.5%), Environmental noise (10721, 5.9%), Music (9169, 5.0%), Silence (6612, 3.6%) and Animal (6216, 3.4%).

B. Sound event detection model

In the next step, a sound event detection neural network model (SED NN) was trained. This model was later used to process the unlabeled field recordings to differentiate between segments that contain and don't contain salient sounds.

Toward this goal, a deep convolutional neural network was designed that consisted of 10 one-dimensional convolutional layers, each followed by batch normalization and ELU nonlinearity. The network was concluded with a flatten layer and an output dense layer. This network accepted on input a vector of 24k raw audio samples (3 seconds at 8kHz) and contained approximately two million trainable parameters.

The network was trained on AudioSet [16], a large-scale dataset of audio clips which was also used for training the PANN model. The training, which used a balanced sample of approximately 100k recordings from AudioSet, was carried out for 150 epochs using Adam [17] optimizer with a learning rate of 0.001, Mean Squared Error loss, and a batch size of 32.

C. Identification of background segments

In this step, we used the SED NN to identify which field recordings' segments were free from salient sounds. The idea of using a model trained on weakly labeled data to perform sound event detection is not new and was extensively explored in [18]. In this context, however, we used the SED NN not to detect any particular types of sounds, but to check if the same types of sounds appear during the whole length of the segment. This was done by moving the relatively short receptive field of the SED NN along the length of each segment and checking if the SED NN predictions are similar.

In particular, we took three-second sections of each segment with a stride of one second and processed them with the SED NN. This yielded eight sections for each segment. For each section, an output from the last convolutional layer of

the SED model was taken (a vector of 128 numbers) as a high-level embedding of that section. The features from the last convolutional layer were used instead of the model's output probability scores in order to achieve an even higher discriminative power for different variations of sounds than provided by the 527 AudioSet classes.

Given the eight embeddings produced by SED NN for each segment, we first calculated an average embedding and then found a maximum Euclidean distance between each embedding and the mean embedding. We observe that if a segment does not contain any salient sounds, this maximum distance should be a relatively small number. Conversely, the maximal distance should be relatively high if a segment contains one or more salient sounds. By setting a threshold on this parameter, the identification of segments without salient sounds should be possible.

The threshold was experimentally set at 2.7 which resulted in 56355 (31%) segments to be classified as not containing salient sounds. Please note that the threshold was deliberately set low to make sure that the selected segments contain no salient sounds.

As a means of subjective validation, a random sample of 2000 background segments was manually inspected by careful visual examination of power spectrograms accompanied by sound playback through headphones. It was found that 1908 segments (95.4%) contained no detectable salient sounds, 70 segments (3.5%) contained faint salient sounds and 22 segments (1.1%) contained clear salient sounds. As such, the accuracy of the proposed background segment identification algorithm was rated high and comparable to the expected accuracy of human raters, who, in a related task, exhibited an accuracy ranging from 81.3% to 95.7% [19].

D. Extraction of salient sounds

In this step, the remaining 126291 segments (69%) were processed using classical signal processing techniques to extract outstanding sounds.

First, the segments were processed with a non-stationary noise reduction algorithm [20] in order to suppress the background. Next, for each segment, the short-term power of the signal was calculated with a sliding window of 250ms. Given these envelopes, individual sounds were extracted from segments by finding optimal cutting points on the time axis which needed to satisfy following conditions: (1) total length does not exceed 4 seconds; (2) maximum power between cutting points is not smaller than -35 bB; (3) power at cutting points is at least 35 dB below maximum power between the cutting points. From each segment, a maximum of 5 sounds were extracted starting from those that had the highest maximum power.

This procedure yielded 57283 sounds extracted from 30494 segments. Again, the parameters for this step were set to get only the clean sounds that stood out from the background in which they were embedded. As a result, 95797 segments (52%) were not used - they were not selected as backgrounds and no salient sounds were extracted from them.

In order to ensure the validity of an approach in which a random salient sound from our dataset is mixed with a random background section, creating a realistically-sounding mixture, a random set of 1000 such mixtures mixed at 0 dB were manually investigated in a similar way as in the validation of background segment selection. It was found that only in 2 of 1000 mixtures (0.2%) the salient sound event was undetectable by a human. This has given a convincing evidence that the size, quality, and diversity of the dataset are sufficient to simulate real-life conditions.

E. Dataset summary and split

The previous steps resulted in a dataset comprising 56355 10-second segments extracted from 24105 field recordings which were found not to contain any salient sounds and 57283 potentially salient sounds (average length 1.5s, minimum 0.3s, maximum 4.0s), extracted from 15412 field recordings. From this dataset, random testing and validation subsets were extracted. This was done by selecting 1500 random original field recordings used in the dataset as a testing subset and additional 1500 original field recordings as a validation subset. The number of background segments and salient sounds in the testing subset was 2442 and 2614, respectively. For the validation subset, these numbers were, respectively, 2509 and 2352. The dataset split was performed based on original field recordings' identifiers from freesound.org to avoid a situation in which segments or sounds from the same original recording are included in different subsets, which would reduce the reliability of the obtained results.

III. Salient sound extraction neural network

Given the dataset of salient sounds, we attempted to train a model capable of detecting and extracting salient sounds from mixtures. For this purpose, we adopted a solution successfully used in the speech enhancement problem: a U-Net neural network trained to predict complex ideal ratio mask (cIRM). The architecture of the proposed model is depicted in Fig 2.

The proposed U-Net consists of a total of 13 convolutional layers organized into six levels. The encoder consists of five downsample modules, each comprising two convolutional layers, from which the second is strided, which realizes the downsampling. Analogically the decoder consists of five upsample modules, each comprising two convolutional layers from which the second is a transposed convolution which realizes the upsampling.

Several optimizations have been introduced in this U-Net which aim at reducing the computational complexity while retaining the quality of predictions. First, the number of filters in all convolutional layers (except for the output layer) is constant and equal to 64, which greatly reduces the number of trainable parameters and processing time. This also allows skip connections to use addition rather than concatenation, which further reduces computational cost. Finally, all batch normalizations and nonlinearities are done after skip connections, which was shown to work very well in residual neural

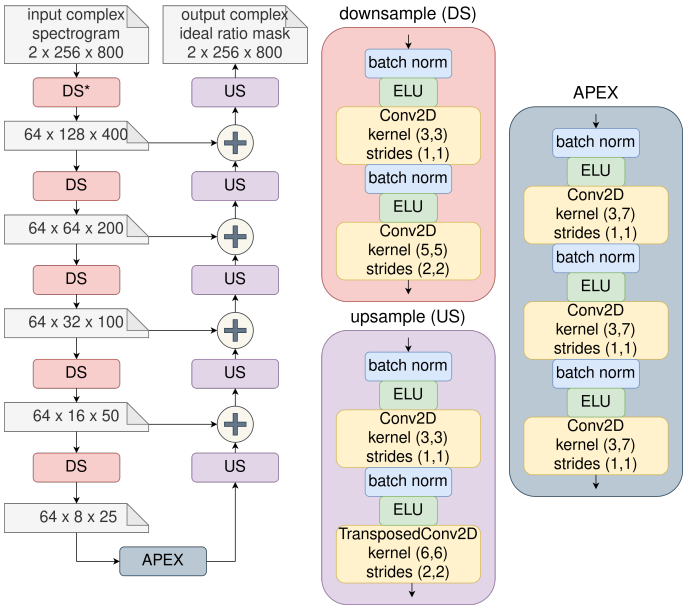


Fig. 2. Architecture of the proposed U-Net for salient sound extraction. First downsample module does not use batch normalization and ELU nonlinearity on input.

networks [21]. The total number of trainable parameters of this model is approximately 1.7M.

The U-Net accepts on input a complex spectrogram calculated using 512-point STFT with Hann window and stride of 10ms. The network estimates complex ideal ratio mask, which, when applied to the input noisy complex spectrogram, extracts the salient sounds by suppressing all background. As suggested in the aforementioned publication, the components of the cIRM were compressed using hyperbolic tangent, mask values were limited to the range of [-10, 10] and the steepness controlling coefficient C was set to 0.1.

The network was trained on the training subset of the dataset described in the previous chapter. The training was conducted for 100 epochs, where in each epoch, each background segment was used once. The preparation of input mixtures included the following steps: (1) randomization of background sound level in the range from -60 to -20 dB; (2) selection of a random 8-second crop from the 10s segment; (3) with a probability of 0.5, selection of a random salient sound and mixing it with the background at a random location on the time axis at the signal-to-noise ratio (SNR) selected uniformly from [-10, -5, 0, 5] dB. As such, half of all examples did not have any salient sounds mixed in. The training mixtures were prepared independently for each epoch, ensuring that the same mixture was never used twice. The mixtures for validation and testing were prepared in a similar manner from respective subsets with the exception that in the case of the testing mixtures the SNRs were predetermined.

The training was performed in batches of 12 mixtures using Mean Squared Error loss. Optimization was performed using Adam [17] optimizer with a learning rate of 0.002 which decayed by 0.97 each epoch. After each epoch, a value of the loss function was checked on the validation set and the model weights with the lowest loss were retained.

IV. Evaluation

The trained model was evaluated on the testing mixtures described in the previous section. The performance of the model was tested in two tasks: (1) detection of the presence of salient sounds, and (2) the ability of an external system to identify the type of salient sound based on the extract.

In the first task, for each of the 2442 test recordings we obtained maximum acoustical energy calculated as a root mean square of samples in a sliding window of 500ms. This value was considered by us as a continuous variable that should differentiate between test examples that have and don't have salient sounds. Indeed, if a mixture doesn't have any salient sounds, the output from our salient sounds extraction system should produce silence (i.e. low, preferably null acoustical energy). Conversely, if a salient sound is present, the output should include such sound therefore the energy should be considerably higher. The efficacy of a model in the detection of the presence of salient sounds was measured using the Area Under the ROC Curve.

In the second task, we once again employed PANN to first obtain a ground truth label for the original salient sound and next to check if the same label was found within top 5 labels returned by the PANN model for the output (processed) recording. The final measure of accuracy was the proportion of test recordings with salient sounds in which the ground truth label was found in top-5 predicted labels (top-5 accuracy). Please note that this scenario is consistent with the expected application of our general-purpose salient sound extraction system as described in the Introduction.

In both tasks three versions of audio files were tested: (1) original (unprocessed) file, (2) file de-noised using non-stationary noise suppression algorithm [20] and (3) salient sound extracted using the proposed U-Net (a result of applying the network-estimated cIRM to the spectrogram of input recording). Additionally, the proposed solution was tested in three variants: (3a) baseline model as described in previous chapters, (3b) model that directly predicts the salient sound as real and imaginary parts of the complex spectrogram (no cIRM), and (3c) a baseline version with additional input channel containing the estimated noise floor of the input mixture.

In the last case, the noise floor was estimated by the Speech Presence Probability Minimum Mean-Square Error (SPP-MMSE) approach with fixed priors [22]. This method assumes Gaussian distributions of the complex coefficients of both noise and speech, and that noise signal is more stationary than speech signal. Using the uniform priors, under which both the speech presence and absence probability are 0.5, the noise floor can be estimated in an MMSE manner. Although originally developed for speech, our initial investigation revealed that the SPP-MMSE noise floor is highly adequate for a wide range of salient sounds present in our dataset. By adding the

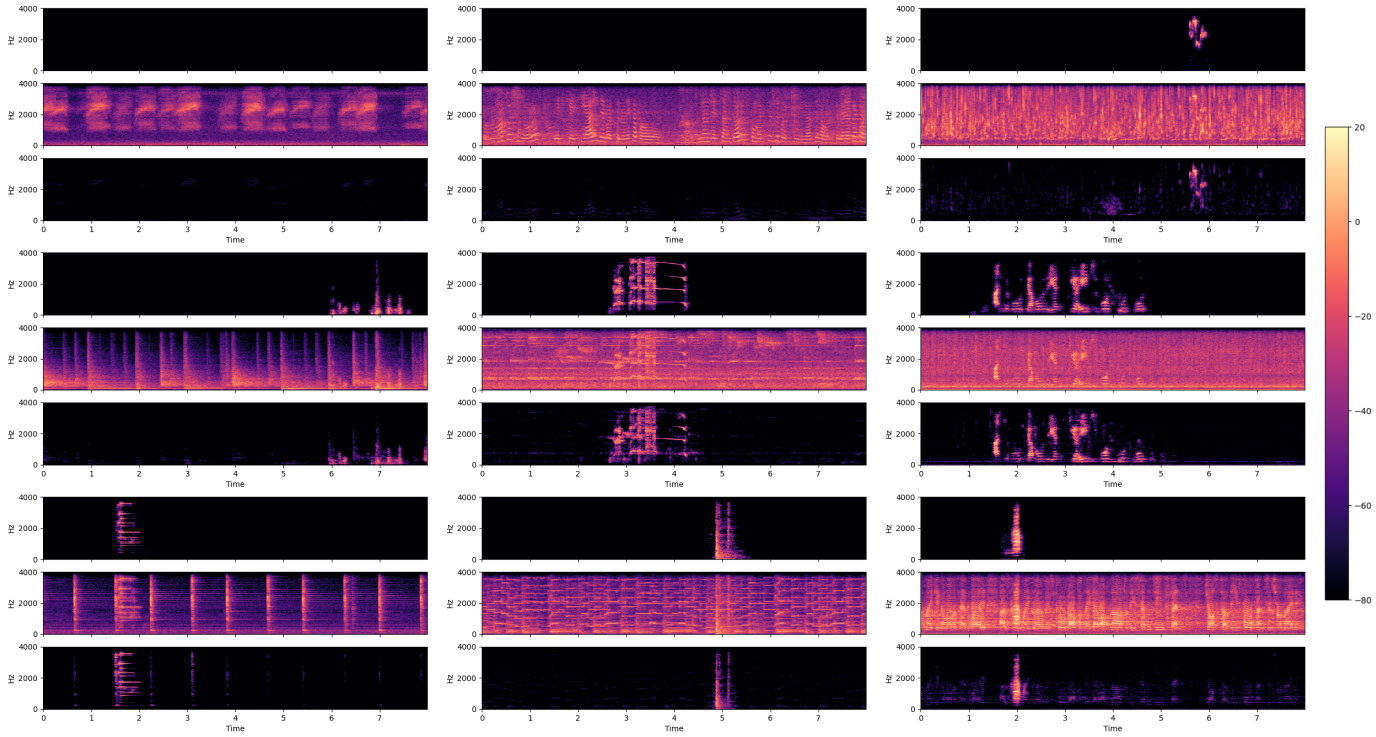


Fig. 3. Nine examples of salient sound extraction using proposed U-Net. Each example is depicted with three power spectrograms: (top) original salient sound; (middle) salient sound mixed with a random background at 0 dB; (bottom) extracted salient sound, obtained by applying the estimated cIRM to the mixture. Each spectrogram shows 8 seconds of audio and the vertical axis is scaled linearly between 0 Hz (bottom) and 4 kHz (top).

noise floor as an additional input channel, we hoped to improve the model's ability to detect and extract these sounds.

All test evaluations were conducted for salient sounds mixed with backgrounds at -10, -5, 0, and 5 dB and additionally a version with pure salient sound (no background) was tested to check how much each solution degrades the pure (background-free) signal.

V. Results

Examples of salient sounds extracted by the proposed U-net are shown in Fig 3. All examples come from the held-out test set. Table I shows the results of the salient sound detection task and Table II shows results of the salient sound type determination experiments.

TABLE I
Efficacy in detecting the presence of salient sound in the processed test recordings using signal level measured by Area Under the ROC Curve (AUC)

Test signal type	Mixture SNRs				Pure salient
	-10 dB	-5 dB	0 dB	5 dB	
1. unprocessed	48.9%	51.8%	59.3%	70.0%	**100%**
2. denoised with [20]	54.8%	65.8%	78.1%	87.9%	**100%**
3a. processed with U-Net - baseline	**85.9%**	**93.4%**	96.9%	98.6%	**100%**
3b. processed with U-Net - direct output	68.7%	82.0%	91.4%	96.1%	99.8%
3c. processed with U-Net - noise floor	85.2%	93.3%	**97.1%**	**98.7%**	**100%**

TABLE II
Top-5 accuracy in determining salient sound type in processed test recordings using an external system (PANN)

Test signal type	Mixture SNRs				Pure salient
	-10 dB	-5 dB	0 dB	5 dB	
1. unprocessed	36.0%	39.3%	43.2%	47.6%	**100%**
2. denoised with [20]	44.9%	47.1%	48.5%	53.7%	91.1%
3a. processed with U-Net - baseline	**66.5%**	**70.6%**	75.4%	**80.2%**	99.2%
3b. processed with U-Net - direct output	53.6%	60.2%	64.2%	68.6%	78.3%
3c. processed with U-Net - noise floor	65.2%	70.4%	**76.0%**	78.7%	98.5%

As can be seen, the proposed solution allows for the detection of the presence of salient sounds with high accuracy even when the sounds are below the noise level. On average, salient sounds were detected with AUC of 93.7%, which is an improvement of 36 percentage points (p.p.) over the unprocessed signals and an improvement of 22 p.p. over the denoised signals.

Determination of the correct type of salient sound is much more challenging, which is partially related to the high sensitivity of the external SED system, PANN. Nevertheless, the proposed solution offered an average accuracy of 73.2%, which is an improvement of 32 p.p. over the unprocessed signals and 25 p.p. over the denoised signals. The results for the pure salient sounds reveal that the proposed solution degrades the salient sounds in a minimal way, much smaller

than the reference noise reduction algorithm.

Additional experiments show that using cIRM as a training target is highly beneficial for the system performance (an average improvement of 11.5 p.p. in salient sound type determination compared to the solution that predicts complex spectrogram of the extracted signal directly). In contrast, adding noise floor as a third input channel to the U-Net affects the performance in a minimal way (an average performance degradation of 0.5 p.p. as compared to the baseline that does not use noise floor on input).

One possible limitation of the evaluated approach for salient sound extraction is the fact that salient sounds, by definition, are rare events, which causes the model training targets to be highly imbalanced. Indeed the values of the expected cIRM masks are, in a vast majority of time-frequency locations, near or equal to zero. This has a negative effect on the neural network model training because, in the initial phases of the training, the models learns to suppress all sounds which already gives a very low value of the loss function. There are multiple ways in which this issue can be alleviated, but their analysis is beyond the scope of this work.

VI. Conclusions

This work presented a novel system for general-purpose salient sound event extraction from monaural audio. To facilitate this, we curated a dataset comprising 56355 audio clips containing salient-free backgrounds and an equally large set of clean salient sounds. These recordings were sourced from a diverse array of unlabeled field recordings. A lightweight U-Net was trained on this data to predict a complex ideal ratio mask. The solution was able to accurately detect the presence of salient sounds in the audio and to isolate them from the background with small distortion, allowing automatic detection of salient sound type by an external system with commendable accuracy. As such, an approach typically used for speech enhancement was found to be effective also in the salient sound detection and extraction problem. However, it was observed that the highly imbalanced nature of the model training targets (a vast majority of time-frequency components of the predicted mask are zeros) might have a negative effect on the training, which indicates a possible directions for future research.

One limitation of the proposed system is the relatively short time window (8 seconds) that is taken into consideration by the model. In many applications, sounds that reappear in an acoustical environment every few seconds should not be considered salient. However, improvement in this regard can be quite easily achieved by repeating the proposed steps for longer segments, depending on the particular demand.

References

[1] Luc H. Arnal, Andreas Kleinschmidt, Laurent Spinelli, Anne-Lise Giraud, and Pierre Mégevand, "The rough sound of salience enhances aversion through neural synchronisation," *Nat. Commun.*, vol. 10, no. 1, pp. 3671, Aug. 2019.

[2] Sandeep Reddy Kothinti and Mounya Elhilali, "Are acoustics enough? semantic effects on auditory salience in natural scenes," *Front. Psychol.*, vol. 14, pp. 1276237, Nov. 2023.

[3] Zied Mnasri, Stefano Rovetta, and Francesco Masulli, "Anomalous sound event detection: A survey of machine learning based methods and applications," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5537–5586, Dec. 2021.

[4] Christoph Kayser, Christopher I. Petkov, Michael Lippert, and Nikos K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, Nov. 2005.

[5] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen, "Sound event detection in the dcase 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, June 2019.

[6] Kevin Wilkinghoff and Fabian Fritz, "On using pre-trained embeddings for detecting anomalous sounds with limited training data," in *2023 31st European Signal Processing Conference (EUSIPCO)*. Sept. 2023, IEEE.

[7] Sania Gul, Muhammad Salman Khan, and Ata Ur-Rehman, "DEW: A wavelet approach of rare sound event detection," *PLoS One*, vol. 19, no. 3, pp. e0300444, Mar. 2024.

[8] Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, A. S. Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A. Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H. Al-Timemy, Ye Duan, Amjed Abdullah, Laith Farhan, Yi Lu, Ashish Gupta, Felix Albu, Amin Abbosh, and Yuantong Gu, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, Apr. 2023.

[9] Tomasz Grzywalski and Szymon Drgas, "Speech enhancement by multiple propagation through the same neural network," *Sensors*, vol. 22, no. 7, 2022.

[10] Yanjue Song and Nilesh Madhu, "Investigations on the optimal estimation of speech envelopes for the two-stage speech enhancement," *Sensors*, vol. 23, no. 14, 2023.

[11] Georgiana-Elena Sfeclis, Ben Milner, and Danny Websdale, "Investigating imaginary mask estimation in complex masking for speech enhancement," in *2023 31st European Signal Processing Conference (EUSIPCO)*. Sept. 2023, IEEE.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[13] Donald S. Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[14] "Freesound — freesound.org," https://freesound.org/, [Accessed 06-07-2022].

[15] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley, "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.

[16] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[17] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.

[18] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019.

[19] Karol J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, New York, NY, USA, 2015, MM '15, p. 1015–1018, Association for Computing Machinery.

[20] Tim Sainburg, "timsainb/noisereduce: v1.0," June 2019.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016, IEEE.

[22] Timo Gerkmann and Richard C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2011.