A-V Tok and Douyin, more audio-visual (With the rise of musical video social platforms such as Tikbedding, cross-modal learning, multi-modal fusion detecting musical video corpus with frame-level labels is introduced. the cross-modal music and audio signal processing, a new live signals is highly beneficial for VAD. To attract more attention to indicating that the incorporation of both audio and visual sig-

modal A-V in the same model framework; 2) the performance of the bi-
networks for VAD was investigated in [9].

A-V results of the branch outperform that of the audio branch in the same model framework; 2) the performance of the bi-

model far outperforms that of audio-only models, indicating that the incorporation of both audio and visual sig-

als is highly beneficial for VAD. To attract more attention to the cross-modal music and audio signal processing, a new live musical video corpus with frame-level labels is introduced.

Index Terms— Audio-visual voice detection, rule embedding, cross-modal learning, multi-modal fusion

1. INTRODUCTION

With the rise of musical video social platforms such as TikTok and Douyin, more audio-visual (A-V) data is uploaded to the Internet. To process these diverse data, voice activity detection (VAD) is an essential preprocessing in detecting the presence or absence of human voice in clips. Classical applications of VAD include speaker diarization [1], music [2] and speech [3] signal processing.

Many VAD methods [4] typically rely on acoustic features of audio signals [5, 6]. These approaches perform well when there is little or no noise in the background. Recently, deep learning has brought more robust methods for VAD. A deep neural network using magnitude and phase information was proposed for VAD under noise condition [7]. Convolutional neural networks have been used as acoustic models for VAD in noisy environments [8]. An optimization-based recurrent neural networks for VAD was investigated in [9].

In live musical streams, an anchor is a person who speaks to the audience or sings in the front of the camera, and plays other music at the same time. This work aims to detect when the anchor speaks or sings in live musical streams, which is a cocktail party-like environment where there may be more than one person singing or speaking at the same time, and the background contains music and other noise. The interference of noise and voice from music in the background easily causes false detection and audio-based VAD may fail.

To make the detector focus more on the target voice and ignore unrelated sounds in the noisy background, a lot of attention has been paid to the audio-visual VAD (AV-VAD). Considering that the lip’s movement is tightly correlated with speech, a statistical VAD model is designed to fuse the audio and lip’s information [10]. By using face parameters (eye and lip contours), a real-time VAD is developed as a front-end for speech recognition [11]. An A-V detector [12] is presented to separate speech from non-speech frames. As mentioned above, previous work on AV-VAD mainly focuses on speech. However, in musical streams, the anchor usually spends more time singing than speaking. The AV-VAD proposed in this paper should detect not only speech, but also the singing of anchor. Due to the different articulation and phonation between speaking and singing, the speech activity detector does not perform well with musical clips [2]. Authors in [13] attempt to detect singing voice and speech based on the same audio-based VAD model, without distinguishing whether the speech or singing voice comes from the anchor or background. Currently, to the best of our knowledge, there is no direct research on singing voice detection using visual information in diverse and noisy scenarios like live musical streams.

This work aims to detect singing voice and speech of the anchor in live musical streams, which is a mixture of polyphonic music and different voices from different people. To better concentrate on the target voice, a simple and effective rule-embedded network is proposed for AV-VAD based on a straightforward observation that visual events usually occur together with acoustic events and they are coordinated [14]. That is, the facial information can directly reflect whether the anchor is vocalizing. This paper explores the possibility of using the facial information as the mask to filter out non-target acoustic events. Experiments show the incorporation of A-V signals is highly beneficial for VAD in noisy scenarios.

Yuabo Hou, Yi Deng, Bilei Zhu, Zejun Ma and Dick Botteldooren

1. Ghent University, Belgium 2. New York University, USA 3. Bytedance AI Lab, China.

1. INTRODUCTION

rule-embedded network for audio-visual voice activity detection in live musical video streams

Yuanbo Hou, Yi Deng, Bilei Zhu, Zejun Ma and Dick Botteldooren

1. Ghent University, Belgium 2. New York University, USA 3. Bytedance AI Lab, China.

ABSTRACT

Detecting anchor’s voice in live musical streams is an important preprocessing step for music and speech signal processing. Existing approaches to voice activity detection (VAD) primarily rely on audio, however, audio-based VAD is difficult to effectively focus on the target voice in noisy environments. This paper proposes a rule-embedded network to fuse the audio-visual (A-V) inputs for better detection of the target voice. The core role of the rule in the model is to coordinate the relation between the bi-modal information and use visual representations as a mask to filter out the information of non-target sound. Experiments show that: 1) with the help of cross-modal fusion using the proposed rule, the detection results of the A-V branch outperform that of the audio branch in the same model framework; 2) the performance of the bi-modal A-V model far outperforms that of audio-only models, indicating that the incorporation of both audio and visual signals is highly beneficial for VAD. To attract more attention to the cross-modal music and audio signal processing, a new live musical video corpus with frame-level labels is introduced.

Index Terms— Audio-visual voice detection, rule embedding, cross-modal learning, multi-modal fusion

1. INTRODUCTION

With the rise of musical video social platforms such as TikTok and Douyin, more audio-visual (A-V) data is uploaded to the Internet. To process these diverse data, voice activity detection (VAD) is an essential preprocessing in detecting the presence or absence of human voice in clips. Classical applications of VAD include speaker diarization [1], music [2] and speech [3] signal processing.

Many VAD methods [4] typically rely on acoustic features of audio signals [5, 6]. These approaches perform well when there is little or no noise in the background. Recently, deep learning has brought more robust methods for VAD. A deep neural network using magnitude and phase information was proposed for VAD under noise condition [7]. Convolutional neural networks have been used as acoustic models for VAD in noisy environments [8]. An optimization-based recurrent neural networks for VAD was investigated in [9].

In live musical streams, an anchor is a person who speaks to the audience or sings in the front of the camera, and plays other music at the same time. This work aims to detect when the anchor speaks or sings in live musical streams, which is a cocktail party-like environment where there may be more than one person singing or speaking at the same time, and the background contains music and other noise. The interference of noise and voice from music in the background easily causes false detection and audio-based VAD may fail.

To make the detector focus more on the target voice and ignore unrelated sounds in the noisy background, a lot of attention has been paid to the audio-visual VAD (AV-VAD). Considering that the lip’s movement is tightly correlated with speech, a statistical VAD model is designed to fuse the audio and lip’s information [10]. By using face parameters (eye and lip contours), a real-time VAD is developed as a front-end for speech recognition [11]. An A-V detector [12] is presented to separate speech from non-speech frames. As mentioned above, previous work on AV-VAD mainly focuses on speech. However, in musical streams, the anchor usually spends more time singing than speaking. The AV-VAD proposed in this paper should detect not only speech, but also the singing of anchor. Due to the different articulation and phonation between speaking and singing, the speech activity detector does not perform well with musical clips [2]. Authors in [13] attempt to detect singing voice and speech based on the same audio-based VAD model, without distinguishing whether the speech or singing voice comes from the anchor or background. Currently, to the best of our knowledge, there is no direct research on singing voice detection using visual information in diverse and noisy scenarios like live musical streams.

This work aims to detect singing voice and speech of the anchor in live musical streams, which is a mixture of polyphonic music and different voices from different people. To better concentrate on the target voice, a simple and effective rule-embedded network is proposed for AV-VAD based on a straightforward observation that visual events usually occur together with acoustic events and they are coordinated [14]. That is, the facial information can directly reflect whether the anchor is vocalizing. This paper explores the possibility of using the facial information as the mask to filter out non-target acoustic events. Experiments show the incorporation of A-V signals is highly beneficial for VAD in noisy scenarios.
The main contributions of this work are: 1) the visual information is introduced into singing voice detection, and unlike previous audio-only VAD studies on either speech or singing voice, this work aims to detect both speech and singing voice using A-V information in live music streams; 2) to fuse the bi-modal inputs simply and reasonably, a rule-embedded network based on the instinctive observation of co-occurrence is proposed; 3) to attract more attention to cross-modal music and audio signal processing, a live musical video corpus is introduced.

This paper is organized as follows, Section 2 introduces the rule-embedded audio-visual multi-branch framework. Section 3 describes the dataset, baseline, experimental setup and analyzes the results. Section 4 gives conclusions.

2. RULE-EMBEDDED AV-VAD NETWORK

The proposed rule-embedded AV-VAD network is shown in Fig. 1. The left part is the audio branch (red words) that tries to learn high-level acoustic features of target events in the audio. The right part is the image branch (blue words) that attempts to judge whether the anchor is vocalizing using visual information. The bottom part is the A-V branch (purple italics), which aims to fuse the bi-modal representations to determine the probability of target events explored in this paper.

2.1. The Audio Branch

The goal of the audio branch is to learn core representations of 4 classes of events: Silence, Speech, Singing, and Others. Speech and Singing contain all speech and singing voice in audio streams, without distinguishing whether the speech or singing voice comes from the anchor or background. The following A-V branch will rely on visual information to determine the source of voices. Others class includes: non-speech and non-singing voices, such as coughing, laughing, cheering and instrumental music, etc.

To learn the acoustic representations, a multi-output convolutional recurrent neural network (CRNN) is used, which is the audio branch connected by red lines in Fig. 1. Waveforms are converted to log mel spectrograms and input into the CRNN. Convolutional layers with gated linear units (GLUs) [15] are applied to learn local shift-invariant patterns from the acoustic feature. By using GLUs, the model can learn to attend to target events and ignore unrelated sounds [16]. The pooling operation is implemented by convolution with strides (1, 2), which means the stride of the convolution along the time is 1 to preserve the time resolution of the input.

To learn the core representations of 4 classes events, 4 outputs with the same structure but different objectives are designed. After the last convolution layer, 4 gated recurrent units (GRU) [17] layers are adopted to capture the temporal context information of Silence, Speech, Singing, and Others, respectively. Following the GRU layers, 4 dense layers are used to obtain the final latent representation of the 4 classes. The outputs of the audio branch are 4 independent units, each of them is a binary classification using a sigmoid to obtain the probability of the corresponding class in the current input. Please visit our homepage for details and source code.

2.2. The Image Branch

The structure of image branch is similar to audio branch, but has a different goal. The image branch aims to predict the possibility of vocalizing and non-vocalizing of the anchor using facial information in the image sequence. Hence, there are 2 outputs with the same structure to learn the core high-level representations of Vocalizing and Non-vocalizing.

To comprehensively consider the continuity of anchor’s actions between the front and rear frames and the contextual information of the image in streams, the input of the image branch is an image sequence of fixed time length. A convolutional structure similar to that in the audio branch is used to extract spatial features of the image sequence. Following convolutional layers, 2 dense layers are applied to extract high-level representations vectors. Like the audio branch, the outputs of the image branch are 2 independent units with sigmoid functions to indicate the probability of the corresponding class in the current input. This structure enables the model to learn the core embedding vector contained in each class.

2.3. The Audio-Visual Branch

Visual events and acoustic events are synchronized [14]. The facial information can reflect whether the anchor is vocalizing intuitively. When the anchor’s mouth is closed, he or she generally does not vocalize; when the anchor’s eyes are closed, he or she may be laughing, yawning, or sleeping instead of

1https://github.com/Yuanbo2020/Audio-Visual-VAD
singing or speaking. Based on the correlation between the anchor’s voice activity and its face parameters, this paper explores the possibility of using the facial information as a mask to filter out non-target acoustic events and focus on the target voice event. The relation among various acoustic events in audio, visual events in image, and target events can be simply represented by Table 1. The main idea of Table 1 is using visual representations as the mask to filter out the information of non-target sound. The relation in Table 1 is simplified to the connection rule embedded in the network in Fig. 1. Note that Speech and Singing in A-V branch only contain the speech and singing voice from the anchor (target speaker), and are different from the Speech and Singing in the audio branch which may also come from the background.

To simply and effectively combine the high-level representations of two branches with different semantics, this paper fuses the embedding vectors according to the rule in Table 1. There are 3 common ways to fuse bi-modal vectors, including Hadmard product (HP) [18], simple concat (SC), and matrix multiplication (MM). The dense layers in the audio and image branch are used to obtain the final latent representation of each class, the corresponding vector is regarded as the abstract feature. Given acoustic and visual embeddings \( \{a, v\} \in \mathbb{R}^{128} \), the operations:

- SC: \( o = \text{Concat}(a, v) \), \( o \in \mathbb{R}^{256} \);
- MM: \( O = av^T, O \in \mathbb{R}^{128 \times 128} \);
- HP: \( o = a \odot v, o \in \mathbb{R}^{128} \);

where the symbol \( \odot \) is the element-wise product. For unified processing, \( o \) and \( O \) will be flattened and input into the corresponding dense layer, as shown in Fig. 1. For bi-modal information, a look up table (LuT) according to Table 1 based on the outputs of the audio and image branch could also be used. The performance of these methods will be comprehensively compared in later experiments.

For training, to consider the cross-modal information at the same time and obtain the corresponding decisions, the audio, image and A-V branch will be trained together. The loss function of the multi-branch AV-VAD model is:

\[
L = \lambda_1 L_{a-sil} + \lambda_2 L_{a-spe} + \lambda_3 L_{a-sin} + \lambda_4 L_{a-oth} + \lambda_5 L_{v-voc} + \lambda_6 L_{v-non-voc} + \lambda_7 L_{av-sil} + \lambda_8 L_{av-spe} + \lambda_9 L_{av-sin} + \lambda_{10} L_{av-oth} \tag{1}
\]

where \( L_a, L_v \) and \( L_{av} \) denote the loss of audio, image and A-V branch; \( sil, spe, sin \) and \( oth \) denote silence, speech, singing and others; \( non-voc \) and \( voc \) denote non-vocalizing and vocalizing. \( \lambda_i \) is the scale factor of each loss function, \( \lambda_i \) determines the importance of each loss function in training.

### Table 1: The relation among acoustic, visual, and target events.

| Audio          | Silence | Speech | Singing | Others |
|----------------|---------|--------|---------|--------|
| Vocalizing     |         | Speech | Singing | Others |
| Non-vocalizing | Silence | Others | Others  | Others |

3. EXPERIMENTS AND RESULTS

3.1. Dataset, Baseline, and Experiments Setup

Since there is no direct research on singing voice detection by A-V information in live streams, and there are fewer A-V datasets that can be used to train the AV-VAD model to detect both target speech and singing voice in noisy environments, a 500-minute musical video dataset (MVD500) with frame-level labels is used in this paper. From this dataset, 360 minutes, 40 minutes, and 100 minutes are used for training, validation, and testing, respectively. To avoid bias caused by gender in training, the total duration of live broadcast of male and female anchors is close in MVD500. To stimulate research in the cross-modal musical signal processing, a 100-minute musical A-V corpus (MAVC100) with frame-level labels is released to our homepage for open research.

As baselines, a common and typical bi-modal recurrent neural model [19] is used as A-V baseline (Base-AV), and a CRNN [2] trained by transfer learning is used as audio-based baseline (Base-A) to compare the performance of the AV-VAD from more perspectives.

In training, log mel-band energy is extracted using STFT with Hamming window length of 44 ms and the overlap is 50% between the window, which has sufficient time-frequency resolution. Then 64 mel filter banks are used. Dropout and normalization are used to prevent over-fitting. Each output in the AV-VAD model is binary classified, and the Adam optimizer [20] is used to minimize the binary cross entropy. Please see our homepage for source code.

For evaluation metrics, event-based precision (P), recall (R), F-score and Error rate (ER) [21] are used. Compared with segment-based metrics used in previous studies [22, 16, 2], event-based metrics are more rigorous and accurate to measure the location of events. Higher P, R, F and lower ER indicate a better performance.

3.2. Results and Analysis

This section tries to answer the following research questions:

- **RQ1:** Which way of cross-modal fusion is the most effective: HP, SC, MM or LuT?

To demonstrate the effect of these fusion methods in the cross-modal learning, the detection results of different fusion ways on the test set of MVD500 are shown in Table 2. Since the activation function of the last classification layer in Base-AV is Softmax, which means the output units are equally important in Base-AV. So, for a fair comparison between the multi-branch model in this paper and Base-AV, all \( \lambda_i \) in \( L \) are 1.

### Table 2: The detection results on the test set of MVD500.

|       | SC | MM | HP | LuT | Base-AV |
|-------|----|----|----|-----|---------|
| ER    | 0.53 | 0.48 | **0.40** | 0.78 | 0.72    |
| P (%) | 75.19 | 78.97 | **84.95** | 62.70 | 66.17   |
| R (%) | 66.54 | 67.43 | **70.77** | 49.25 | 53.41   |
| F-score (%) | 70.60 | 72.75 | **77.22** | 55.17 | 59.11   |


In Table 2, the cross-modal fusion by HP performs best, followed by MM, while SC performs the worst. The reason may be that HP is a fuzzy variant of an AND operation, based on HP, the visual embedding vector functions as a mask to help the model attend to target voice and ignore unrelated sound in the noisy background. Each bit in the high-level acoustic feature vector is adjusted via the corresponding visual information, the audio representations that are beneficial to the task are retained and irrelevant content is reduced. SC retains the information without any filtering or enhancement effect, and does not guide the exploration of model. The output of MM has richer information such as the correlation between different bits of audio and image vector, but these information are not independent of each other. Because \( \text{rank}(O) \) is 1, the output matrix of MM is rank deficient. And after flattening, the 128×128-dimensional output of MM brings a great burden to the calculation of the model.

All results of rule-based combination (SC, MM and HP) are better than the Base-AV. This may be because the audio and image branch of Base-AV are just a common feature extractor, and not optimized for the task. The network structure in this paper is specifically designed for AV-VAD, so it can learn the specific high-level features of target events, rather than the universal representation of all events. The results of LuT also show the two branches in the AV-VAD network are effective, because LuT does not involve cross-modal fusion, and only combines the results of two branches according to Table 1. The result of LuT is not much worse than that of Base-AV, indicating that the output of two branches did learn the knowledge of corresponding events as we expected.

• RQ2: How much the A-V branch improves the results compared with that of the audio branch in the same bi-modal model? That is, if the bi-modal inputs and combination rule are useful, using visual information as the mask to filter out irrelevant acoustic representations in the audio, how much gain can this operation bring to the final result?

The gain of using visual information in the cross-modal model is shown in Table 3. Both AV-VAD and Base-AV achieve better results than the audio branch alone, which shows that both models integrate the A-V information well, so that the result of the fusion of different modal information outperforms that of mono-modal information. In detail, the gain of AV-VAD on the F-score and \( \text{ER} \) is 7.73% and 0.16, while that of Base-AV is 2.23% and 0.03. AV-VAD is not only better on the audio branch, but its corresponding visual information gain is also better than that of Base-AV.

• RQ3: Compared with audio-based VAD, how much improvement is obtained by AV-VAD?

The ER and F-score of the audio-based Base-A is 0.86 and 40.93%, respectively. And the corresponding values of AV-VAD shown in Table 3 are 0.40 and 77.22%, respectively. The performance of bi-modal model far outperforms that of audio-based Base-A, which means the rule-embedded AV-VAD is more effective in noisy audio-visual environment. A intuitive detection result of AV-VAD from the live broadcast is shown in Fig. 2. Please see our homepage for video demos.

The above experiments are all tested on the test set of MVD500. To make it easier for others to compare our method and show results more comprehensively, the detection results of AV-VAD on the public MAVC100 is shown in Table 4.

| Class        | Event-based | Segment-based |
|--------------|-------------|---------------|
|              | ER P (%) R (%) F (%) | ER P (%) R (%) F (%) |
| Singing      | 0.48 38.36 49.59 70.11 | 0.13 92.65 95.37 93.11 |
| Speech       | 0.36 90.65 70.78 79.49 | 0.21 97.01 80.85 88.20 |
| Overall      | 0.44 80.83 70.40 75.26 | 0.15 95.35 86.93 90.94 |

4. CONCLUSION

This paper proposes a rule-embedded network to fuse the cross-modal information for audio-visual VAD in live musical video streams. Experiments show the proposed cross-modal fusion method and multi-branch network are effective, and the performance of bi-modal model far outperforms that of an audio-only model. Besides, to attract more attention to the multi-modal music and audio signal processing, a new live musical video corpus with frame-level labels is introduced.

5. ACKNOWLEDGEMENTS

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.
6. REFERENCES

[1] Y. Huang, T. Otsuka, and H. G. Okuno, “A speaker diarization system with robust speaker localization and voice activity detection,” in Contemporary Challenges and Solutions in Applied Artificial Intelligence, pp. 77–82. Springer, 2013.

[2] Y. Hou, F. K. Soong, J. Luan, and S. Li, “Transfer Learning for Improving Singing-Voice Detection in Polyphonic Instrumental Music,” in Proc. Interspeech 2020, pp. 1236–1240.

[3] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” Speech communication, vol. 52, no. 1, pp. 12–40, 2010.

[4] K. Lee, K. Choi, and J. Nam, “Revisiting singing voice detection: A quantitative review and the future outlook,” in In Proceedings of the International Society for Music Information Retrieval (ISMIR), 2018, pp. 506–513.

[5] A. L. Berenzweig and D. P. Ellis, “Locating singing voice segments within music signals,” in Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics. IEEE, 2001, pp. 119–122.

[6] L. Lu and H. Zhang, “Automated extraction of music snippets,” in Proceedings of the eleventh ACM international conference on Multimedia, 2003, pp. 140–147.

[7] L. Wang, K. Phapatanaburi, Z. Go, S. Nakagawa, M. Iwahashi, and J. Dang, “Phase aware deep neural network for noise robust voice activity detection,” in 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017, pp. 1087–1092.

[8] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, “Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 2519–2523.

[9] G. Gelly and J. Gauvain, “Optimization of rnn-based speech activity detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 3, pp. 646–656, 2017.

[10] P. Liu and Z. Wang, “Audio-visual voice activity detection,” Frontiers of Electrical and Electronic Engineering in China, vol. 1, no. 4, pp. 425–430, 2006.

[11] C. T. Ishi, M. Sato, N. Hagita, and S. Lao, “Real-time audio-visual voice activity detection for speech recognition in noisy environments,” in Auditory-Visual Speech Processing, 2010.

[12] D. Dov, R. Talmon, and I. Cohen, “Audio-visual voice activity detection using diffusion maps,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 4, pp. 732–745, 2015.

[13] T. Kosaka, I. Suga, and M. Inoue, “Improving voice activity detection for multimodal movie dialogue corpus,” in 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE). IEEE, 2018, pp. 481–484.

[14] A. Senocak, T. Oh, J. Kim, M. Yang, and So K., “Learning to localize sound source in visual scenes,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4358–4366.

[15] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in Proceedings of the 34th International Conference on Machine Learning - Volume 70, 2017, pp. 933–941.

[16] Y. Hou, Q. Kong, S. Li, and M. D. Plumbley, “Sound event detection with sequentially labelled data based on connectionist temporal classification and unsupervised clustering,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 46–50.

[17] R. Dey and F. M. Salemt, “Gate-variants of gated recurrent unit (gru) neural networks,” in IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE, 2017, pp. 1597–1600.

[18] R. A. Horn, “The hadamard product,” in Proc. Symp. Appl. Math, 1990, vol. 40, pp. 87–169.

[19] F. Tao and C. Busso, “End-to-end audiovisual speech activity detection with bimodal recurrent neural models,” Speech Communication, vol. 113, pp. 25–35, 2019.

[20] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in The 3rd International Conference for Learning Representations, San Diego, 2015.

[21] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” Applied Sciences, vol. 6, no. 6, pp. 162, 2016.

[22] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, “Sound event detection in the DCASE 2017 challenge,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.