CATEGORY-ADAPTED SOUND EVENT ENHANCEMENT WITH WEAKLY LABELED DATA

Guangwei Li, Xuenan Xu, Heinrich Dinkel[1]*, Mengyue Wu*[†]*, Kai Yu*[†]

MoE Key Lab of Artificial Intelligence
X-LANCE Lab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai, China
[1]Xiaomi Corporation, Beijing, China

ABSTRACT

Previous audio enhancement training usually requires clean signals with additive noises; hence commonly focuses on speech enhancement, where clean speech is easy to access. This paper goes beyond a broader sound event enhancement by using a weakly supervised approach via sound event detection (SED) to approximate the location and presence of a specific sound event. We propose a category-adapted system to enable enhancement on any selected sound category, where we first familiarize the model to all common sound classes and followed by a category-specific fine-tune procedure to enhance the targeted sound class. Evaluation is conducted on ten common sound classes, with a comparison to traditional and weakly supervised enhancement methods. Results indicate an average 2.86 dB SDR increase, with more significant improvement on speech (9.15 dB), music (5.01 dB), and typewriter (3.68 dB) under SNR of 0 dB. All enhancement metrics outperform previous weakly supervised methods and achieve comparable results to the state-of-the-art method that requires clean signals.

Index Terms— source separation, weakly supervised learning, deep neural networks, category adaptation

1. INTRODUCTION

Audio enhancement is an important front-end to improve performance for back-end tasks such as automatic speech recognition (ASR), speaker identification (SI), and music classification. Traditionally, training and testing of an enhancement system require clean sources of targeted sounds. Take speech enhancement as an example. Standard training procedure involves a mixture of clean speech and other additive noises. Recent advances in deep learning promote the deployment of neural networks as computational backbones in audio enhancement [1, 2, 3, 4, 5, 6, 7]. While neural networks can achieve state-of-the-art (SOTA) performance for audio enhancement and source separation tasks [8, 9], they are generally trained with a supervised training objective thus are limited to clean source signals.

A few works start to explore unsupervised methods [10, 11] or embed weakly-labeled data for source separation and speech enhancement [12, 13, 14]. [10, 11] use permutation invariant training (PIT) to estimate each composition of the original wave mixture. [12, 13] select segments from weakly-labeled data and use a conditional vector to control the source to be separated or enhanced. Though providing an illuminating training method and an encouraging result, enhancement and separation with no given clean signals under real-world scenarios is still an emerging field. For instance, weakly-labeled datasets involve various real-world sound types while the current enhancement is limited to speech, music, and few other categories.

Therefore, this paper proposes an enhancement system on any targeted category with weakly-labeled real-world data. Compared with previous weakly-supervised enhancement approaches, which are trained from scratch, our system is trained by a *generalized separation - adapted enhancement* paradigm, including two stages. The model is first pre-trained to separate general sound events and then fine-tuned to enhance a targeted sound class. The two-stage training paradigm familiarizes the model to all common sound classes before enhancement training, leading to better overall enhancement performance. We select ten commonly-seen sound events from the original 527 sound classes in Audioset, including Speech, Music, Vehicle, Animal, Train, Guitar, Dog, Horse, Alarm, and Typewriter. We evaluate the enhancement and generalization performance of our method on these classes to validate the effectiveness and robustness of our category-adapted enhancement system under a signal-to-noise ratio (SNR) of 0 dB. As for our speech-adapted enhancement system, we further compare it with other traditional and weakly supervised enhancement methods. We also present a few enhanced examples as demos for a more intuitive presentation.

2. CATEGORY-ADAPTED ENHANCEMENT

The proposed system pipeline is shown in Figure 1, consisting of three primary procedures: (1) Anchor segments selecting and categorizing, where a sound event detection (SED) model is used to select valid segments from an audio clip of a given class and estimate the probabilities of the sound events; (2) A generalized 527-categories separation system by randomly choosing a wave pair from two sound categories; (3) A category-adapted enhancement system, realized by fine-tuning the separation system with re-selected anchor segments of the targeted sound class using a stricter boundary estimation process.

2.1. Anchor segments selecting and categorizing

Since frame-level annotation is unavailable in weakly labeled datasets, we use an SED system to obtain frame-level predictions, based on which we select anchor segments for effective training. The SED model is a convolutional recurrent neural network (CRNN), as described in Section 3. We first select anchor segments based

[†] Mengyue Wu and Kai Yu are the corresponding authors.

This work has been supported by National Natural Science Foundation of China (No.61901265), State Key Laboratory of Media Convergence Production Technology and Systems Project (No.SKLMCPTS2020003) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

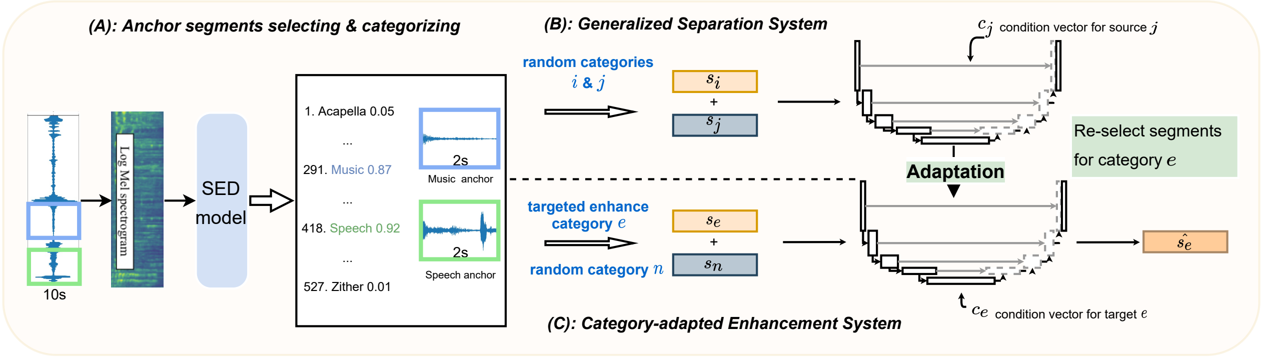
(A): Anchor segments selecting & categorizing
10s
Log Mel spectrogram
SED model
1. Acapella 0.05
...
291. Music 0.87
...
418. Speech 0.92
...
527. Zither 0.01
2s
Music anchor
2s
Speech anchor
(B): Generalized Separation System
random categories i & j
s_i
+
s_j
c_j condition vector for source j
Re-select segments for category e
Adaptation
targeted enhance category e
random category n
s_e
+
s_n
$\hat{s_e}$
c_e condition vector for target e
(C): Category-adapted Enhancement System

Fig. 1. The proposed category-adapted enhancement system architecture. The pipeline first selects anchor segments by using a pre-trained SED model, categorized by their respective event label (A). Then a generalized separation system using randomly selected pairs of sound events is trained (B). Lastly, the generalized system is adapted towards a target sound class (C).

on the frame-level SED output with the highest probability. That is, for a specific audio clip w_k, with the offered sound events label k, the expected presence probability of sound class k is denoted as $O_k(t) \in [0, 1], t = 1, 2, ..., T$ where T represents the output length of the SED model.

The time step τ with the largest probability of sound class k is used to select an anchor segment s_k with duration τ_0:

$$\begin{aligned} \tau &= \arg\max_t O_k(t) \\ s_k &= w_k[\tau - \frac{1}{2}\tau_0, \tau + \frac{1}{2}\tau_0] \end{aligned} \tag{1}$$

In this way, we can ensure the existence of anchor segments from all available sound classes (e.g., 527 in Audioset), enabling the training of a generalized separation system.

2.2. Generalized separation system

The proposed separation model is based on the U-Net structure [15], which is fed a spectrogram of mixed sources and outputs the separated spectra. We use the anchor segments from 527 classes to train this generalized separation system.

We choose two segments s_i and s_j from two random sound classes i and j. Usually, only one clean sound class is used for training a separation system due to data limitation, while the segments we choose in Section 2.1 often contain more than one sound class. A conditional vector c_k is thus used to control which source is to be separated. The separation system can be described as:

$$f(s_i + s_j, c_k) \rightarrow s_k \tag{2}$$

where $k = i, j$. We set $c_k \in [0, 1]^K$ to represent the presence probabilities of all K sound classes. The n-th number of $c_k, n \in 1, 2, ..., K$ is calculated from the frame-level SED output $O_k(t)$, using linear softmax [16]:

$$c_k[n] = \frac{\sum_t^T O_k[n]^2(t)}{\sum_t^T O_k[n](t)} \tag{3}$$

When choosing the wave pairs for training, the prerequisite is that the two segments include different sound classes.

$$c_i \cdot c_j \geq \eta \tag{4}$$

Given a resemblance threshold η, we reject any combination of two segments $s_{i,j}$ if the dot product of the condition vectors c_i and c_j meets Equation (4). In other words, we assure that the separation system is trained on diverse sound event pairs.

2.3. Category-adapted enhancement system

After training the generalized separation system, we further aim to enhance the separation performance by adapting to a pre-defined category. This adaptation process fine-tunes the generalized separation system by changing the training objective to focus on separating a target category from other sound types.

Thus, we adapt the system to a category-specified enhancement system with a segment re-selecting procedure. Compared with training the enhancement system from the very beginning, our separation-enhancement adaptation enables the model to learn features of all available sound classes so that it can recognize and eliminate otiose sound classes. Under this paradigm, it is possible to train a robust, any-category enhancement system from the pre-trained generalized separation system.

2.3.1. Re-selecting segments for the targeted class

When training the separation system, involving anchor segments from all sound classes aims to enhance the model's generalization ability. During the category-adapted enhancement stage, a different segment selecting strategy is needed for a more specified, targeted sound class enhancement.

Anchor segments are re-selected with a stricter rule. Take speech enhancement as an example, for audio clips with a referenced speech tag, we use the double-threshold [17] method to find a continuous audio chunk (longer than 2 s) containing speech, shown in Figure 2. The first column is an example of three 10-second audio clips with selected 2-second speech segments. With the double-threshold method, the first segment is slightly revised while the third is object to a major time-shift, and the second segment is discarded. Our approach differs from previous work in [13], which verified the validity of a sound event within a window of 320 ms. High quality segments of the targeted sound class will improve the performance of the enhancement system to be trained.

2.3.2. Enhancement system training

We use the proposed category-adaptation procedure to train an enhancement system for any given class. Suppose we are to train an

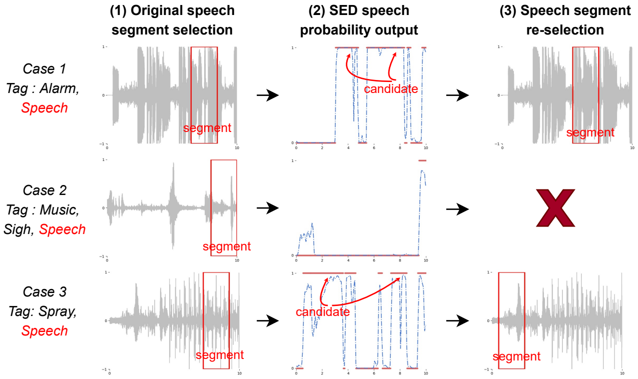
(1) Original speech segment selection
(2) SED speech probability output
(3) Speech segment re-selection
Case 1
Tag : Alarm, Speech
segment
candidate
segment
Case 2
Tag : Music, Sigh, Speech
segment
Case 3
Tag: Spray, Speech
segment
candidate
segment

Fig. 2. The re-selection algorithm for targeted wave segments (e.g., speech). For each case, the red bounding box in the first column marks the original segment chosen. The second column shows the speech probability (blue dot-dash) and continuous region (red line) given by the SED model. The red box in the third column marks the re-selected speech segment. Case 1,2 and 3 show the three different situations where re-selected segment is unchanged, discarded and largely time-shifted, respectively.

enhancement system for sound class e, the separation model (Section 2.2) and the re-selected e class wave segments (Section 2.3.1) are incorporated. We mix e segments with another category n, screened with Equation (4) to ensure that the sound events in these two segments are different. Training is done by learning the following objectives:

$$f(s_e + s_n, c_e) \rightarrow s_e \tag{5}$$

$$f(s_e, c_e) \rightarrow s_e \tag{6}$$

$$f(s_e, c_n) \rightarrow \mathbf{0} \tag{7}$$

For both separation (Equation (2)) and enhancement (Equations (5) to (7)) training, the input source signal s is transformed to a spectrogram denoted as S using short-time Fourier transform (STFT). The model is trained to minimize the mean square error (MSE) between the expected spectrogram magnitude $|\tilde{S}|$ and the estimated spectrogram magnitude $|\hat{S}|$:

$$\mathcal{L}_{\text{MSE}} = \left\| |\tilde{S}| - |\hat{S}| \right\|_2 \tag{8}$$

During inference, the estimated magnitude $|\hat{S}|$ and the phase of the input source $\angle S$ are used to recover the spectrogram of the estimated signal $\hat{S} = |\hat{S}|e^{j\angle S}$.

3. EXPERIMENTS

While our proposed system can enhance any class within Audioset, we pick 10 commonly-seen sound classes including Speech, Music, Vehicle, Animal, Train, Guitar, Dog, Horse, Alarm and Typewriter. We focus on validating the effectiveness of our approach on these classes.

Enhancement data preprocessing We use the *balanced* training subset (21,155 audio clips) from Audioset, consisting of 527 sound event classes, to train and evaluate our separation and enhancement system. Each audio clip is weakly labeled and includes one or more sound classes. We evaluate our approach in regards to ten sound types mentioned above. They are mixed with natural noises from another noise sound type respectively as out test inputs. Speech is from the clean test-set of Librispeech [18] and contains 2620 clean speech utterances. Music is taken from the MUSAN dataset [19] (total duration ≈ 42 h) while noise is from the free sound category in the MUSAN noise dataset (≈ 6 h). Other categories of sound are from the FindSounds, an audio search engine containing a huge amount of sound clips of different categories.

Experimental setup We use an SED model to predict the onset and offset and give out the predicted presence of a given sound class in an audio clip. Our SED model is trained on the unbalanced (≈ $5000h$) subset of Audioset [20], consisting of 527 sound events. The SED model is a convolutional recurrent neural network (CRNN) named L-CDur with eight convolution blocks, which is based on CDur [21]. Each block contains a convolution layer, a batch normalization layer, and a ReLU activation. The architecture uses mean-max pooling [17] as its subsampling strategy, and dropout is utilized between convolutional blocks. A log-mel spectrogram (LMS) is fed to obtain the respective expectation $O_k(t)$ for each sound class k at a resolution of 20 ms. The resemblance threshold η is set to 0.4 when selecting the segment pairs.

Following previous work [12], a U-Net is used for the generalized separation and adapted enhancement system. The input segments have a length of 2 seconds and a sample rate of 16000 Hz. STFT features are extracted every 16ms with a window size of 64ms. The U-Net consists of 4 encoders and 4 decoders, whereas the condition vector c_i is added through a learnable linear layer after each convolutional layer. The generalized separation training in the first stage takes 200k iterations and then the model is fine-tuned for category-adapted enhancement. Adam optimizer is used for training the separation and enhancement system with a starting learning rate of 0.001.

During inference, the conditional vector is set to a one-hot vector. An inverse STFT (ISTFT) is used to restore the estimated source from the output spectrogram. The framework is implemented using PyTorch [22].

Metrics Common enhancement metrics are applied: sound-distortion ratio (SDR). For speech enhancement, perceptual evaluation of speech quality (PESQ), and the short-time objective intelligibility (STOI) are also applied.

4. RESULTS

4.1. Enhancement performance for the chosen 10 categories

The results using our proposed category adapted separation system are shown in Table 1.

The first row shows the SDR of the original noisy input. The second row shows the enhancement result of the traditional wiener filter [24, 23], which uses spectral subtraction for enhancing the signal. It has a relatively low performance for most sound classes. The third row is the result of a comparable separation model from [12], which is also trained on the balanced subset in Audioset. The fourth row displays the result of our proposed method. Our category-adapted enhancement surpasses the previous weakly labeled method (3rd row) in almost all sound classes and achieves an average 3.061 dB over the chosen ten sound classes.

We apply our proposed enhancement approach on several examples, including audio clips in the test set, real 10s-videos from the

Table 1. Enhancement performance (SDR) over the chosen 10 sound classes under the sound-noise-ratio (SNR) of 0 dB.

Configuration	Sound classes										Average
	Speech	Music	Animal	Vehicle	Dog	Guitar	Train	Horse	Typewriter	Alarm	
Noisy	0.144	0.139	0.188	0.144	0.219	0.185	0.332	0.164	0.224	0.228	0.197
Wiener filter [23]	2.414	-4.479	-0.609	-8.331	**2.984**	-4.579	-5.335	0.505	0.955	-0.34	-1.682
Kong *et al.* [12] (Baseline)	7.209	1.521	-2.391	1.489	-1.571	1.569	**1.628**	-0.332	2.496	-1.49	1.013
Category-adapted (Proposed)	**9.291**	**5.147**	**1.609**	**2.785**	1.351	**3.031**	1.042	**0.724**	**3.906**	**1.723**	**3.061**

Audioset evaluation set, and movie trailers on YouTube. The demos are available online[1]. All of the enhanced audio and videos are presented without any post-processing.

4.2. Detailed results for Speech-adapted enhancement

Among the ten sound classes chosen above, Speech enhancement is the most popular in audio and signal processing. Many researchers work on this topic and have proposed various methods. Table 2 displays the detailed adapted-enhancement results on speech. Our proposed category-adapted enhancement system achieves 9.291 dB, 2.408, 0.837 in SDR, PESQ, and STOI respectively. Results are analysed by comparing with traditional & state-of-the-art enhancement methods and previous work using similar weakly-supervised pipelines.

Table 2. Speech enhancement performance of different configurations under 0 dB SNR noise. *Adapted-n* = our proposed training scheme, an speech-adapted enhancement from n steps of generalized separation training.

Configuration	Enhancement		
	SDR ↑	PESQ ↑	STOI ↑
(1) Clean	\	\	\
(2) Noisy	0.144	1.899	0.794
(3) Wiener filter [23]	2.414	1.680	0.750
(4) Tasnet [8, 9]	**10.372**	**2.573**	**0.871**
Other weakly-supervised methods			
(5) Kong Separation [12]	7.209	2.080	0.815
(6) Kong Enhancement [13]	8.496	2.063	0.804
(7) Proposed (Adapted-200k)	**9.291**	**2.408**	**0.837**
(8) w/o two-stage	7.081	2.100	0.809
(9) Adapted-100k	8.890	2.314	0.826
(10) Adapted-300k	9.261	2.356	0.825

Comparison with other enhancement approaches Wiener filtering [23] is a traditional signal processing method often deployed in denoising. Applying a wiener filter has an adverse effect on intelligibility performance, leading to lower PSEQ and STOI scores, see Line (3) in Table 2.

Modern neural network speech enhancement approaches such as Tasnet [8, 9], the current SOTA, require clean labeled data. We use a Tasnet trained on WSJ0 [25] and Wham [26] as a topline comparison. It achieves a superior enhancement performance, reaching an SDR of 10.372 dB and 2.573, 0.871 for PESQ and STOI respectively. However, as a weakly supervised method, our method approaches SOTA without using clean speech training data.

Comparison with other weakly-supervised approaches As mentioned above, weakly-labeled data on Audioset has been utilized for separation and speech enhancement training in previous works [12, 13]. Our proposed *generalized separation - adapted enhancement* paradigm (7) outperforms these weakly supervised approaches, achieving an SDR improvement of 0.795 dB. With the exclusion of the two-stage training paradigm (i.e. training the enhancement system from scratch (8)), the performance degrades significantly (SDR 9.291 → 7.081), indicating the effectiveness of the proposed paradigm.

We make further ablations by changing the training iterations of the separation model before adapted to category-enhancement (9) (10). Both the increase and decrease of pre-training iterations lead to performance drop. This indicates the importance of balance between our first stage separation and second stage enhancement. Training from the separation model helps the model learn more about various sound features besides the targeted category. However, if the first stage is completely trained, the enhancement performance of our category-adapted method begins to decrease.

Visualization Visualized examples of speech enhancement are also provided (see Figure 3), suggesting the generalized effectiveness of our adapted-enhancement paradigm.

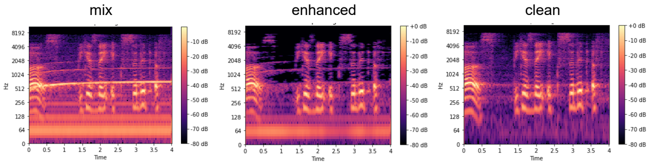

Fig. 3. Power spectrograms of mix, enhanced and clean speech.

5. CONCLUSION

In this paper, we propose a category-adapted enhancement system with weakly labeled data using a two-stage *generalized separation - adapted enhancement* training paradigm. It achieves an average SDR improvement of 3.06 dB over the ten chosen sound classes. Specially, our speech-adapted enhancement system achieves an average of 9.291 dB, 2.408, and 0.837 in SDR, PESQ, and STOI under SNR of 0 dB noise. The results indicate that our proposed approach achieves significant performance improvement over traditional and other weakly supervised approaches in heavy noise scenarios on most sound classes. Our speech-adapted enhancement approaches the performance of the supervised training paradigm without the requirement of clean speech signals.

[1] https://ligw1998.github.io/source-separation.html

6. REFERENCES

[1] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[2] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proceedings of Conference of the International Speech Communication Association*, 2013.

[3] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International conference on latent variable analysis and signal separation*. Springer, 2015, pp. 91–99.

[4] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proceedings of Conference of the International Speech Communication Association*, 2017, pp. 3642–3646.

[5] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[6] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang, N. Kamo, M. Hira, T. Hayashi, C. Boeddeker, Z. Chen *et al.*, "Espnet-se: end-to-end speech enhancement and separation toolkit designed for asr integration," in *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 785–792.

[7] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[8] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[9] ——, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, vol. 33, 2020, pp. 3846–3857.

[11] ——, "Unsupervised speech separation using mixtures of mixtures," *Workshop on International Conference on Machine Learning (ICML)*, 2020.

[12] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source separation with weakly labelled data: An approach to computational auditory scene analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 101–105.

[13] Q. Kong, H. Liu, X. Du, L. Chen, R. Xia, and Y. Wang, "Speech enhancement with weakly labelled data from audioset," in *Proceedings of Conference of the International Speech Communication Association*, 2021, pp. 191–195.

[14] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proceedings of Conference of the International Speech Communication Association*, 2020.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[17] H. Dinkel and K. Yu, "Duration Robust Weakly Supervised Sound Event Detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2019, pp. 311–315.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[19] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[21] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 887–900, 2021.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2019, vol. 32, pp. 8026–8037.

[23] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[24] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[25] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop*, 1992.

[26] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proceedings of Conference of the International Speech Communication Association*, Sep. 2019.