Listen to What You Want: Neural Network-based Universal Sound Selector

Tsubasa Ochiai, Marc Delcroix, Yuma Koizumi, Hiroaki Ito, Keisuke Kinoshita, Shoko Araki

NTT Corporation, Japan

tsubasa.ochiai.ah@hco.ntt.co.jp

Abstract

Being able to control the acoustic events (AEs) to which we want to listen would allow the development of more controllable hearable devices. This paper addresses the AE sound selection (or removal) problems, that we define as the extraction (or suppression) of all the sounds that belong to one or multiple desired AE classes. Although this problem could be addressed with a combination of source separation followed by AE classification, this is a sub-optimal way of solving the problem. Moreover, source separation usually requires knowing the maximum number of sources, which may not be practical when dealing with AEs. In this paper, we propose instead a universal sound selection neural network that enables to directly select AE sounds from a mixture given user-specified target AE classes. The proposed framework can be explicitly optimized to simultaneously select sounds from multiple desired AE classes, independently of the number of sources in the mixture. We experimentally show that the proposed method achieves promising AE sound selection performance and could be generalized to mixtures with a number of sources that are unseen during training.

Index Terms: sound extraction, deep learning, acoustic event

1. Introduction

In our daily lives, we are constantly immersed in acoustic scenes or environments that are composed of mixtures of acoustic events (AEs), which are often called polyphonic sounds [1, 2]. Depending on the situation, AEs can provide critical information about our surroundings, e.g., klaxons when crossing a street. They can also be annoying disruptions of our concentration, e.g., similar klaxons when working at home with an open window. Our daily lives could be greatly improved if we could develop hearable devices that select the AEs that we want to listen to depending on the situation. Toward this goal, we define two problems: 1) AE sound selection and 2) AE sound removal. The AE sound selection problem consists of the extraction of the one or multiple desired AE sounds from the mixture, given the user-specified target AE classes. When multiple desired AE classes are selected, we output a signal that consists of the sum of all the AEs from these classes. In this paper, we focus on the AE sound selection problem, and briefly discuss the related AE sound removal problem, which consists of the suppression of the undesired AE sounds from the mixture. To the best of our knowledge, this paper is the first attempt that explicitly tackles these two problems.

Thus far, research on AE processing has focused on AE detection/classification (AED) and AE separation (AES): separating a mixture of AEs into individual AEs. AED detects the boundaries and identifies the classes of each AE in a recording, which may have many practical applications, e.g., for surveillance [3, 4]. AEDs have been extensively researched, and rapid progress in the field has been fueled by the recent DCASE challenge series [5] and the release of large datasets [6, 7]. Since AED enables only to detect sounds without extracting them

from the acoustic scene, they are thus insufficient to solve the AE selection problem that we target.

Recently, the separation of AEs using neural networks to output signals for each AE in a mixture has received increased interest [8, 9]. For example, the utterance-level permutation invariant training (uPIT) [10] framework, which was initially developed for speech processing, has been extended for the separation of AEs [8]. Although such an approach could separate the AEs, it has two main drawbacks. First, the number of AEs that it can separate is fixed by the number of network outputs. In the case of speech separation, it would be reasonable to fix the maximum number of outputs to two or three since in practice it is rare that more relevant speakers speak within an audio segment. However, the number of AEs can be substantially larger in an acoustic scene. Training a PIT-based separation network for a larger number of outputs is challenging because of the exponential increase in the number of permutation possibilities in the PIT loss. Another issue is the global permutation ambiguity at the separation's output, i.e., it is ambiguous which output corresponds to which source AE. Therefore, AES alone also fails to solve the AE selection problem.

One way to achieve AE sound selection would be to combine AES and AED. For example, with a cascade combination of AES and AED obtained by performing AED after a PIT based AES [11], the desired AEs could be picked up from the outputs. This approach has three limitations. First, it inherits the drawbacks of PIT-based separation in terms of the maximum number of AEs it can manage. Second, the AED performance after separation may not be optimal because of the possible processing artifacts or remaining undesired sounds in the separation output. Finally, we need to perform AED on all the separation outputs, which may become computationally expensive when dealing with many possible simultaneous AEs.

In this paper, we propose a neural network-based AE sound selection approach, called Sound Selector, which directly extracts the desired AE sound from a mixture of AEs given a one-hot vector representing the class of interest. Since the Sound Selector outputs a single signal independently of the number of AE sounds in the mixture, it does not require knowing the maximum number of AEs like a PIT-based AES. Moreover, the Sound Selector is explicitly optimized based on an objective function for AE sound selection. Finally, we extend the Sound Selector to simultaneously extract AEs from multiple AE classes, making it computationally efficient, which would be important for future extensions of the approach to real-time processing on hearable devices.

2. Proposed method

In this paper, we define two novel concepts, i.e., AE sound selection and AE sound removal. In the following for discussion simplicity, we focus on AE sound selection and describe our proposed framework that extracts (selects) the desired AE sounds from the observed mixture (Figure 2 for an example), given the user-specified target AE classes.
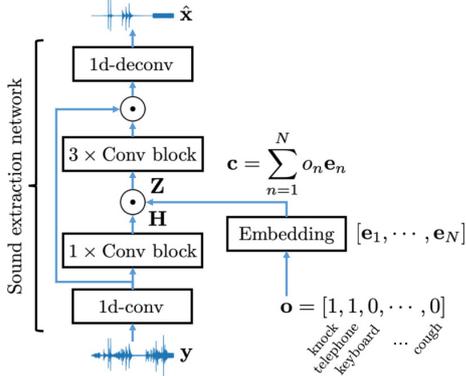
Figure 1: *Overview of proposed Sound Selector architecture.*

2.1. Sound selection framework

Let $\mathbf{y} \in \mathbb{R}^T$ be the T-length time-domain waveform of the observed mixture. We assume that target-class vector $\mathbf{o} = [o_1, \cdots, o_N]^\mathsf{T}$ is also given as input, where N denotes the number of total AE classes. For example, to extract the AE sounds of the n-th AE class, $\mathbf{o}$ is a one-hot vector, where the element corresponding to the target AE class is set to $o_n = 1$ and others are set to $o_j = 0 \;\; \forall j \neq n$. Given observed mixture $\mathbf{y}$ and target-class vector $\mathbf{o}$ as inputs, the time-domain waveform of target AE class $\hat{\mathbf{x}}$ is directly estimated by the sound extraction network, as follows:

$$\hat{\mathbf{x}} = \mathrm{DNN}(\mathbf{y}, \mathbf{o}), \qquad (1)$$

where $\mathrm{DNN}(\cdot)$ is the non-linear transformation of a deep neural network (DNN). Motivated by the success of time-domain speech separation/extraction frameworks [12, 13], we adopt Conv-TasNet architecture, which demonstrated remarkable speech separation performance. Thus, $\mathrm{DNN}(\cdot)$ mainly consists of stacked dilated convolution blocks.

Figure 1 shows a schematic diagram of the network architecture of our proposed AE sound selector network, i.e., a Sound Selector, which consists of an AE-class embedding layer and a sound extraction network. Given $\mathbf{y}$ as input, the bottom block of the sound extraction network generates an intermediate representation of observed mixture $\mathbf{H} = \{\mathbf{h}_1, \cdots, \mathbf{h}_F\}$, where $\mathbf{h}_f \in \mathbb{R}^{D \times 1}$ denotes the feature at f-th frame, F is the total number of frames and D is the dimension of the feature space. In parallel, given $\mathbf{o} \in \mathbb{R}^{N \times 1}$ as input, an AE-class embedding layer generates target-class embedding $\mathbf{c} \in \mathbb{R}^{D \times 1}$, which provides an encoded representation of the target AE class. These intermediate representations $\mathbf{H}$ and $\mathbf{c}$ are combined into an integrated representation $\mathbf{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_F\}$, which is passed to the upper blocks of the sound extraction network to output only the sounds from the target AE class. By integrating the target-class embeddings $\mathbf{c}$ into the sound extraction network, the network behavior can be adapted to extract the target AE sounds.

There are several ways that could be used to perform the above integration procedure. In this paper, motivated by the success in speaker adaptation [14] and target speech extraction [15], we adopt an elementwise product-based integration, such as $\mathbf{z}_f = \mathbf{h}_f \odot \mathbf{c} \;\; \forall f$.

2.2. Extension to multi-class simultaneous extraction

We above described AE sound selection for a single AE class. However, the Sound Selector is required to extract arbitrary sounds desired by a user depending on the situation. This can include sounds from multiple AE classes.

To extract the sounds of multiple AE classes, we could iteratively apply the extraction module for all the target classes and then sum up the extracted signals to reconstruct the desired output. However, such an iterative approach naturally increases the computational cost as the number of target class increases.

To alleviate the above issue, we propose a multi-class simultaneous extraction scheme that extracts multi-class AE sounds in a one-pass manner and makes the computational cost constant as the number of target classes increases. In a multi-class selection setting, the network output remains a single signal, which includes all the sounds belonging to the target classes.

In this work, we hypothesize that the network behavior can be adapted to extract not only a single target AE class but also an arbitrary combination of multiple target AE classes. To inform the sound extraction network about the multiple AE classes, we compute target-class embedding $\mathbf{c}$, as follows:

$$\mathbf{c} = \mathbf{E}\mathbf{o} = \sum_{n=1}^{N} o_n \mathbf{e}_n, \qquad (2)$$

where $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_n, \cdots, \mathbf{e}_N]$ is a trainable embedding matrix and $\mathbf{e}_n \in \mathbb{R}^{D \times 1}$ represents the embedding vector of the n-th AE class. $\mathbf{E}$ is jointly optimized with the other parts of the network. This formalization corresponds to having a target-class vector $\mathbf{o}$ set to a n-hot vector, where the n elements that correspond to the target AE classes are 1 and the others are 0.

2.3. Training procedure

We assume that a set of input and target features $\{\mathbf{y}, \mathbf{o}, \{\mathbf{x}_n\}_{n=1}^N\}$ is available for training the model, where $\mathbf{x}_n \in \mathbb{R}^T$ is the target sound signal of the n-th AE class. In general, the number of AE sounds in the mixture is lower than the number of total AE classes N. We thus set the AE sounds of classes $\mathbf{x}_n$ that do not exist in the mixture at zero signals.

To realize the proposed multi-class simultaneous extraction, we dynamically generated target-class vector $\mathbf{o}$, where one or more elements are 1 and the others are 0, and also generated the corresponding reference signal, as follows:

$$\mathbf{x} = \sum_{n=1}^{N} o_n \mathbf{x}_n. \qquad (3)$$

To retain the scale information of the AE sounds, we adopted the scale-dependent signal-to-noise ratio (SNR) [16] as the training objective. The SNR loss $\mathcal{L}$ is expressed as follows:

$$\mathcal{L} = 10 \log_{10}\left(\frac{\|\mathbf{x}\|^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|^2}\right) \Leftrightarrow -10 \log_{10}\left(\|\mathbf{x} - \hat{\mathbf{x}}\|^2\right), \quad (4)$$

where $\hat{\mathbf{x}}$ denotes the estimate of the target signal computed from $\mathbf{y}$ and $\mathbf{o}$. For the baseline PIT-based separation, we use the equivalent log mean squared error (MSE) criterion to enable training when the targets are zero signals.

2.4. Sound removal problem

The sound removal problem is the opposite of the sound selection problem. Sound selection aims to extract the sounds of the selected AE classes, while sound removal aims to suppress the sounds of the selected AE classes from the observed mixture.

We consider two variants to realize a sound removal mechanism: 1) direct estimation and 2) indirect estimation. In the former scheme, we build a sound removal network by changing the reference signal in Eq (3) to removal target $\mathbf{x} = \mathbf{y} -$

$\sum_{n=1}^{N} o_n \mathbf{x}_n$. In the latter scheme, we use the Sound Selector to extract the sounds to suppress and generate the removal output as $\hat{\mathbf{x}} = \mathbf{y} - \hat{\mathbf{x}}^{\text{Sel.}}$, where $\hat{\mathbf{x}}^{\text{Sel.}}$ denotes the estimate by the Sound Selector.

3. Related work

We base our proposed AE Sound Selector on target speech extraction framework [15, 17], which outputs only a target speech given the target speaker's query. Recently, a similar query-based extraction framework was also proposed in music source separation task [18]. The advantage of this framework is that it can directly optimize the source extraction objective without requiring knowledge of the maximum number of sources in the mixture. In this work, we introduce learned class embeddings instead of source embeddings that are derived from sound examples [15, 17, 18], which allow the extraction of AEs depending on user-specified target classes. We also extend the approach to simultaneously extract AEs from multiple classes, which enables to directly optimize it for the extraction of multiple user-specified target classes, unlike [15, 17, 18] that are optimized for only a single target.

Combination of AES and AED has been investigated for improving AED performance for polyphonic sounds [19, 20]. For example, in recent DCASE challenges, non-negative matrix factorization (NMF)-based separation [21, 22] was often used as a front-end for AED, e.g., [23, 24]. Besides, [9] investigated an AES and AED combination to improve separation performance but not to realize AE sound selection. It derived an iterative PIT-based separation framework, that uses the AED class posteriors of all the separated signals as auxiliary inputs to a second separation pass. This approach inherits the limitations of the PIT-based separation frameworks, i.e., where the maximum number of sources in the mixture must be fixed.

Quite recently, [25] investigated using dataset with weak-labels for AES. In concurrent and independent work, they proposed a similar framework for extracting sounds in a mixture based on condition vector derived from an AED and evaluated their system on mixtures of 2 classes. Compared to [25], we focus on the AE sound selection/removal and investigate explicit extraction of multiple AE classes simultaneously for both training and inference. Besides, we evaluate our proposed scheme for extraction of up to four AE classes simultaneously on mixtures of up to seven AE classes. In future work, we will include an investigation of using weak-labels for training [25, 26].

4. Experiments

4.1. Dataset

To evaluate the effectiveness of our proposed method, we created datasets of simulated sound event mixtures based on the Freesound Dataset Kaggle 2018 corpus (FSD) [6], which contains audio clips from 41 diverse AE classes, such as human sounds, object sounds, musical instruments, etc [6]. We included stationary background noise to the mixtures using noise samples from the REVERB challenge corpus (REVERB) [27]. We created two datasets: 1) three-class mixtures (Mix 3) and 2) three to five-class mixtures (Mix 3-5). The mixtures in Mix 3 contain the AEs of three classes within an audio clip, and the mixtures in Mix 3-5 contain the AEs of three, four, or five classes.

We generated six-second mixtures by randomly extracting six audio clips of 1.5 to 3 seconds from the FSD corpus and pasting (adding) them to random time-positions on top of the six-second background noise. With this way of constructing the data, different sounds from the same AE class can occur

Table 1: *Baseline SDR [dB] for mixture signals.*

Dataset	# class for Sel.	# class in Mix.			
		3	4	5	mean
Mix 3	1	-3.6	-	-	-3.6
Mix 3-5	1	-3.3	-5.9	-7.2	-6.0
	2	3.5	0.0	-2.1	-0.2
	3	21.8	5.9	2.0	6.3

up to twice per mixture. We created sound event mixtures by utilizing Scaper's functionality [28], "generate_from_jams(·)", which is widely used to create dataset in the SED community, where *ref_db* was set at -50 and *snr* was set randomly between 15 and 25 dB for each foreground event. In this experiment, we downsampled the sounds to 8 kHz to reduce the computational and memory costs.

The training and development sets consist of 50,000 and 10,000 mixtures, respectively. The audio clips for these sets were randomly selected from the *training* sets in the FSD and REVERB corpora. The test set consists of 10,000 mixtures based on the *test* set in the FSD and REVERB corpora. For each audio clip, we also randomly generated target-class vector $\mathbf{o}$ that represents the desired AE classes.

4.2. Configurations

For all the experiments, we adopted the non-causal Conv-TasNet-based network architecture, which consists of stacked dilated convolution blocks. By following the notations of [12], we set the hyperparameters as follows: N = 256, L = 20, B = 256, H = 512, P = 3, X = 8, and R = 4. We also set the dimension of embedding layer D at 256. For the integration layer, we adopted element-wise product-based integration and inserted it after the first stacked convolution block (Figure 1).

We adopted the Adam algorithm [29] for optimization with an initial learning rate of 0.0005 and used gradient clipping [30]. We stopped the training procedure after 200 epochs.

As the evaluation metrics, we used the scale-invariant signal-to-distortion ratio (SDR) of BSSEval [31]. In the experiment, we conducted three types of evaluations for 1) single-class selection, and multi-class selection with 2) two classes, and 3) three classes. For each mixture, three AE classes $\{n_1, n_2, n_3\}$ were pre-defined. For the multi-class (I-class) selection task, the reference signal for the SDR calculation is given by $\mathbf{x} = \sum_{i=1}^{I} \mathbf{x}_{n_i}$, where I is the number of target AE classes, i.e., $I \in \{1, 2, 3\}$ in this experiment.

Table 1 shows the average SDR scores of the mixture signals, which are used as the baseline to compute the SDR improvement scores. Here, "# class for Sel." denotes the number of AE classes, i.e., I, defined as the selection target, and "# class for in Mix." denotes the subsets of the evaluation dataset with three, four, or five AE classes in the mixture.

4.3. Results

4.3.1. Evaluation: single-class selection

In the first experiment, we evaluated the single-class extraction task on Mix 3 and Mix 3-5 tasks. Table 2 shows the SDR improvement of the PIT-based system and our proposed Sound Selector. For a fair comparison, we employed a similar network architecture for both the PIT-based and our proposed systems. With the PIT-based system, we assumed oracle target class selection (OS), which selects one of the output sources that shows the highest SDR score as the target AE sound. The PIT-based networks have three and five output channels for the Mix 3 and Mix 3-5 tasks, respectively. They are trained to output zero sig-

Table 2: *SDR improvement [dB] for single-class selection.*

| Method | Dataset | # class in Mix. | | | |
		3	4	5	mean
PIT + OS	Mix 3	12.0	-	-	12.0
Sound Selector	Mix 3	11.2	-	-	11.2
PIT + OS[1]	Mix 3-5	9.6	9.1	8.5	9.0
Sound Selector	Mix 3-5	8.8	10.5	11.1	10.5

Table 3: *SDR improvement [dB] for multi-class selection.*

| Method | # class for Sel. | # class in Mix. | | | |
		3	4	5	mean
Iterative	2	4.4	6.7	8.0	6.9
Simultaneous	2	5.1	6.8	7.7	6.9
Iterative	3	-11.0	2.7	5.2	2.0
Simultaneous	3	-4.9	3.9	5.0	3.3

nals when the number of sources in the mixture is fewer than the output channels [32].

Table 2 shows that the SDR score of the Sound Selector is slightly lower than the PIT+OS for Mix 3 task, but the proposed Sound Selector stably worked better than the PIT+OS for a more realistic Mix 3-5 task, i.e., where the number of classes in the mixture is variable. These results demonstrate that the proposed architecture can jointly solve the AES and AED tasks.

Note that the scores of the PIT+OS would correspond to the upper bound performance with the oracle AE classifier. In practice, performing AED on top of the separation would be difficult due to potential distortion or remaining undesired sounds in the signals[2].Therefore, if we adopt the actual AE classifier, the scores of the PIT+OS will certainly be degraded. Evaluation with such a real classifier would be important future work.

4.3.2. Evaluation: multi-class selection

In the second experiment, we evaluated the multi-class extraction task on the Mix 3-5 task. Table 3 shows the SDR improvement of the iterative and simultaneous extraction schemes. Here, "Simultaneous" denotes the proposed multi-class simultaneous extraction scheme, as described in Section 2.2. The "Iterative" scheme iteratively applies single-class extraction for each of the target AE classes and sums up all of the extracted sounds.

Table 3 shows that the "Simultaneous" extraction scheme works comparably or better than the "Iterative" scheme, regardless of its lower computational cost. These results suggest that the proposed AE selection concept could be achievable with a constant computational cost.

We observe that the SDR improvement is negative when both "# class for Sel." and "# class for Mix." are three. This is due to the high baseline SDR (21.8 dB in Table 1), because the reference signals closely resemble the mixtures in this setting. Note that for the "Simultaneous" scheme, although SDR is degraded by 4.9 dB, the absolute SDR remains high (16.9 dB), showing that the output signal is not significantly distorted.

Due to space limitations, we are unable to report the detailed results on the sound removal experiments, but we confirmed that both direct and indirect estimation schemes per-

---

[1]We observed that PIT-based training with SNR loss failed to output the correct number of sources in the mixture when trained on the Mix 3-5 set. Additional investigations with alternative training objectives or curriculum learning may be required to prevent this issue.

[2]Even for clean signals, the MAP@3 score is about 95 % on the same FSD corpus [6] and the top-1 accuracy would be even lower.
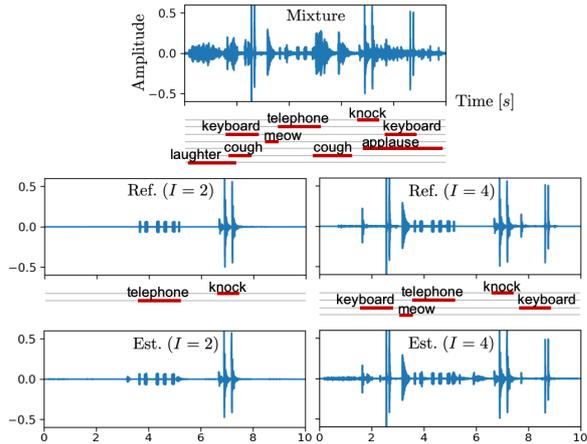

Figure 2: *Mixture with AEs from 7 AE classes, the reference signals (Ref.), and the signals estimated with the proposed Sound Selector (Est.) for 2 (left) and 4 (right) target AE classes.*

formed equivalently with SDR improvement of about 6 dBs.

4.3.3. Analysis of generalization capability

Finally, we explored the generalization capability of our approach to conditions unseen during training, i.e., longer mixtures with more AE classes and more target classes for selection. We created an additional test set consisting of 200 home office-like mixtures of ten seconds containing AEs from seven classes. The target AE classes are set for all mixtures to "knock, telephone" for two-class ($I = 2$) and "knock, telephone, keyboard, meow" for four-class ($I = 4$) cases.

Figure 2 shows an example of sound selection for a AE mixture[3], where "Ref" and "Est" denote reference and estimated signals, respectively. The results were obtained with the multi-class simultaneous extraction scheme. From the figure, we confirmed that our proposed system successfully extracts the sounds of the multiple target AE classes, even if seven-class mixtures and four-class simultaneous extraction were not included in the training stage.

Figure 2 shows only one example, but we observed that the average SDR improvements on this set are 8.5 dB for two-class and 5.3 dB for four-class cases. This result suggests that the proposed framework has the potential to generalize to unseen conditions.

5. Conclusions

In this paper, we introduced two novel concepts, i.e., AE sound selection and removal problems. Solving these problems would open a path toward hearable devices that allows the selection of AEs to which we want to listen. We proposed a sound selection neural network, i.e., Sound Selector, which exploits a n-hot representation of the user-specified target AE classes to simultaneously output sounds from these classes. Experimental results showed that the proposed Sound Selector successfully extracts the multiple AE sounds simultaneously and it has the potential to generalize to an unseen number of classes in the mixture and to the extraction of an arbitrary number of classes.

Future work will include investigations using a larger dataset with more AE classes, which would be essential for further evaluating the generalization capability of the proposed method, as well as extension to online processing and training with weak-labels.

---

[3]Audio examples of the proposed sound selection/removal system will be available online [33]

6. References

[1] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *International joint conference on neural networks (IJCNN)*, 2015, pp. 1–7.

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

[3] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo*, 2005, pp. 1306–1309.

[4] P. Atrey, N. Maddage, and M. Kankanhalli, "Audio based event detection for multimedia surveillance," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006, pp. 813–816.

[5] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 Challenge setup: Tasks, datasets and baseline system," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[6] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2018.

[7] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[8] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 175–179.

[9] E. Tzinis, S. Wisdom, A. Jansen, J. R. Hershey, A. J. Ellis, and P. Daniel, "Improving universal sound separation using sound classification presentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 96–100.

[10] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[11] http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments (accessed on May 15th, 2020).

[12] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 8, pp. 1256–1266, 2019.

[13] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 691–695.

[14] L. Samarakoon and K. C. Sim, "Subspace LHUC for fast adaptation of deep neural network acoustic models." in *Interspeech*, 2016, pp. 1593–1597.

[15] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Cernocky, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[16] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[17] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Interspeech*, 2017, pp. 2655–2659.

[18] J. H. Lee, H.-S. Choi, and K. Lee, "Audio query-based music source separation," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[19] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Machine Listening in Multisource Environments*, 2011.

[20] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste *et al.*, "An exemplar-based nmf approach to audio event detection," in *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, 2013, pp. 1–4.

[21] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[22] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 414–421.

[23] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[24] K. M. Jeon, H. Kim, and H. K. Kim, "Nonnegative matrix factorization-based source separation with online noise learning for detection of rare sound events," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[25] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 101–105.

[26] F. Pishdadian, G. Wichern, and J. Le Roux, "Learning to separate sounds from weakly labeled scenes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 91–95.

[27] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[28] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[30] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2013, pp. 1310–1318.

[31] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing (TASLP)*, vol. 14, no. 4, pp. 1462–1469, 2006.

[32] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5739–5743.

[33] http://www.kecl.ntt.co.jp/icl/signal/member/tochiai/demos/universal_sound_selector/index.html.