MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Learning to Separate Sounds From Weakly Labeled Scenes

Pishdadian, Fatemeh; Wichern, Gordon; Le Roux, Jonathan

TR2020-038 April 11, 2020

Abstract

Deep learning models for monaural audio source separation are typically trained on large collections of isolated sources, which may not be available in domains such as environmental monitoring. We propose objective functions and network architectures that enable training a source separation system with weak labels. In contrast with strong time-frequency (TF) labels, weak labels only indicate the time periods where different sources are active in this scenario. We train a separator that outputs a TF mask for each type of sound event, using a classifier to pool label estimates across frequency. Our objective function requires the classifier applied to a separated source to output weak labels for the class corresponding to that source and zeros for all other classes. The objective function also enforces that the separated sources sum to the mixture. We benchmark performance using synthetic mixtures of overlapping sound events recorded in urban environments. Compared to training on mixtures and their isolated sources, our model still achieves significant SDR improvement.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

LEARNING TO SEPARATE SOUNDS FROM WEAKLY LABELED SCENES

Fatemeh Pishdadian[1,2], *Gordon Wichern*[1], *Jonathan Le Roux,*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Interactive Audio Lab, Northwestern University, Evanston, IL, USA

fpishdadian@u.northwestern.edu, {wichern, leroux}@merl.com

ABSTRACT

Deep learning models for monaural audio source separation are typically trained on large collections of isolated sources, which may not be available in domains such as environmental monitoring. We propose objective functions and network architectures that enable training a source separation system with weak labels. In contrast with strong time-frequency (TF) labels, weak labels only indicate the time periods where different sources are active in this scenario. We train a separator that outputs a TF mask for each type of sound event, using a classifier to pool label estimates across frequency. Our objective function requires the classifier applied to a separated source to output weak labels for the class corresponding to that source and zeros for all other classes. The objective function also enforces that the separated sources sum to the mixture. We benchmark performance using synthetic mixtures of overlapping sound events recorded in urban environments. Compared to training on mixtures and their isolated sources, our model still achieves significant SDR improvement.

Index Terms— audio source separation, semi-supervised classification, weakly-labeled data

1. INTRODUCTION

Supervised methods using deep neural networks have recently demonstrated state of the art performance in speech enhancement [1, 2], speech separation [3, 4], music separation [5–8], and sound effect separation [9]. These approaches typically require a large dataset of isolated sources to generate sound mixtures and their corresponding training targets.

When isolated sources are not available, it is unrealistic for humans to manually label the audio at the granularity of a time-frequency (TF) bin, especially to do so accurately and at scale. It is, however, reasonable to assume they can produce limited labels for the activity and type of sounds within some time range [10, 11]. The annotation burden can be further reduced, as done in sound event detection (SED) [12–18], by replacing the fine resolution labels on the sound event onsets and offsets (e.g., on the order of 10 ms) by a coarse temporal label indicating the presence or absence of a sound event within an audio clip (e.g., on the order of 10 s). Since the fine resolution labels are typically defined at the level of a short-time Fourier transform (STFT) frame, we refer to them as frame-level labels, while we refer to the coarse labels as clip-level labels.

The goal of this paper is to consider whether deep learning separation methods that are typically trained in a supervised way using the TF-bin level labels, can be trained using weaker labels such as frame- or clip-level labels. The use of weakly-labeled data, for the aforementioned practical reasons, has been extensively researched for the SED task. This task is particularly important in our work since we not only try to address similar problems in transitioning from strong to weak labels, but also employ an SED mechanism as the critic for the separation performance.

We shall first point out an important difference regarding the notion of strength of a label depending on the task. In SED, the goal is to estimate the type of an audio event together with its precise onset and offset. As such, the frame- and clip-level labels are respectively referred to as strong and weak labels. In contrast, in the context of source separation, ground truth consists in having information on each source at the TF-bin granularity. Strong labels for SED are thus only weak labels for source separation. There are also key differences in the type of pooling required in SED and source separation. In weakly-labeled SED, consecutive time frames often share the same class labels; in weakly-labeled separation, the structure is much more intricate, as frequency bins sharing the same label may be far from each other, often harmonically spaced in a highly variable manner even among the same types of sounds.

To tackle these difficulties in pooling over the frequency dimension, we propose a form of discriminative pooling, where an SED classifier is employed as the principal metric for loss calculation while training the separator. When applied to a separated source, the classifier is expected to detect that only a single class is present, while all other sources are inactive. We further propose a multi-task learning approach in training the separator, combining the audio event classification objective with an additional separation-specific objective that enforces the separated sources to sum to the mixture. Our model learns to perform separation based solely on weak labels, either at the frame level or at the clip level.

Several works have attempted to train source separation networks with relaxed training data requirements, such as semi-supervised methods [19, 20] which do not require the isolated sources to match the mixture, or those seeing only a target source combined with background, and isolated backgrounds [21, 22]. Another class of methods based on weak labels assumes the availability of weak labels at both training and inference time such as the score-informed approach in [23], the variational auto-encoder in [24], and the audio-visual approach in [25], where the video provides (weak) labels to guide the audio separation. Our approach can separate multiple source classes, does not require seeing any sources in isolation, and requires only the audio mixture (no labels) during inference.

Another line of research performs source separation implicitly when training SED systems [26,27]. The objective function in [27] is only SED cross-entropy and does not include any terms modeling the separation task explicitly, such as enforcing each separated source to belong to a single class, or enforcing estimated sources to sum to the mixture as in our approach. Moreover, they test their method only on isolated sources in background noise, whereas our experiments deal with multiple overlapping sound events.

This work was performed while F. Pishdadian was an intern at MERL.

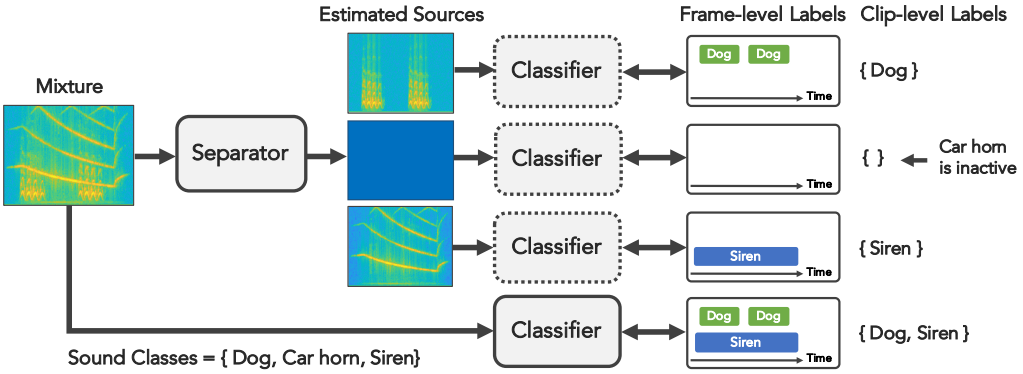

Fig. 1. The joint separation-classification model. The separator receives an audio mixture and returns source estimates. The classifier alternately processes the mixture and estimated sources (dashed lines indicate shared parameters). The classifier is pre-trained on the mixtures to return the presence probabilities for all classes. The separator is then trained such that the classifier applied on a source estimate returns the presence probability for that source along with zeros for all other sources.

2. PROPOSED METHOD

In this section, we present our joint separation-classification approach to audio source separation using weak labels. We describe our model and formulate the objective function.

2.1. Joint Separation-Classification Model

In this work, we address the under-determined source separation problem and assume only one recording channel of the mixture is available. A common approach in this scenario is to perform masking on the mixture in the time-frequency (TF) domain. Let $X_{\omega,\tau}$ denote the mixture magnitude in the TF domain (e.g., magnitude STFT), and $\hat{M}_{i,\omega,\tau}$ a TF mask estimate for the i-th source, taking values in $[0, 1]$. The masking operation can be formulated as

$$\hat{S}_{i,\omega,\tau} = \hat{M}_{i,\omega,\tau} X_{\omega,\tau}, \tag{1}$$

where $\hat{S}_{i,\tau,\omega}$ is the estimated magnitude of the i-th source. The time-domain source estimates are then obtained by applying an inverse transform (e.g., iSTFT) to the estimated magnitudes combined with the mixture phase. Throughout this work, we assume each *sound source* in the mixture belongs to a distinct *sound class* (e.g., speech, music, gun shot), and hence we use these terms interchangeably.

Supervised mask inference refers to training a model to generate mask estimates for all sources present in an audio mixture. In *fully-supervised* mask inference, the TF-domain representations of the isolated sources or TF masks built from them are used as targets in training the model. We refer to such targets as "strong labels", as they provide information about sound classes at the TF-bin level. In the *weakly-supervised* case, however, the TF-bin labels are not available. Instead, the target labels only indicate the presence or absence of sound classes over some duration of time (e.g., in a 32 ms time frame or a 4 s audio clip).

At a high level, our model is composed of two main blocks: a source separator and a source classifier. The block diagram of the entire system is shown in Fig. 1. The separator block receives a TF representation of a mixture and outputs estimates $\hat{\mathbf{S}}_i$, $i = 1, \ldots, n$, for each of the sources, where n indicates the total number of sound classes available in a dataset. We assume the number of active sound classes in a given mixture ranges from 1 to n. The input to the classifier block is also a TF representation. In general, the TF representation used as input to the classifier may be different from the one used as input to the separator, as long as we can pass gradients through the transform used to compute it. For instance, the classifier input can be a mel spectrogram while the separator input is a magnitude STFT.

Our main idea is that, if we can train a classifier that performs well in predicting source class activities on natural mixtures, where sound classes may sometimes occur in isolation and other times overlap with other classes, we may use that classifier as a critic of a separator's performance. We may thus use weak labels, either at the frame or clip level, to train the separator through the classifier. This is illustrated in Fig. 1, where we have shown both frame-level labels, with onsets and offsets for each sound class, and clip-level labels where only presence or absence within a clip is indicated.

The classifier can be trained independently or jointly with the separator. However, training the separator requires the classifier output, as TF-bin labels are not available. We here pre-train the classifier on the set of mixtures and fix its weights when training the separator. Note that the classifier is not trained using the set of isolated sources as targets as this would violate the assumption that strong labels are not available, although some sections with isolated sources may naturally occur in the mixtures.

2.2. Objective Function

Our main goal in training the model is to achieve high-quality separation, which requires explicit optimization of mask estimates, even if ground truth TF labels are not available. To this end, a key constraint is to enforce the output signals of the separator to add up to explain the input mixture. Indeed, this term is critical in preventing the separator from producing masks that only focus on the most discriminating TF components for classification without fully reconstructing the entire source. This is formulated as a mixture magnitude loss $\mathcal{L}_{\text{mix}}(\tau)$ that minimizes the discrepancy between the mixture magnitude and the sum of estimated source magnitudes at each time frame τ:

$$\mathcal{L}_{\text{mix}}(\tau) = \sum_{\omega} |X_{\omega,\tau} - \sum_{i=1}^{n} \hat{S}_{i,\omega,\tau}|, \tag{2}$$

where $|.|$ denotes the modulus operator. Using the information provided by weak labels, we can enforce that only the sum over active sources should be equal to the mixture, and all inactive sources should be silent. Moreover, we can locate mixture frames where no sources are active and exclude them from loss computation. In our experiments, these refinements to the separation loss proved very important for obtaining good mask estimates.

The classifier should correctly identify the sound classes, whether it is applied to the input mixture or any of the source estimates. This can be achieved by including a binary cross-entropy term between the classifier output for each source and the true source label. Let $H(l,p) = -l\log(p) - (1-l)\log(1-p)$ be the binary cross-entropy function where $l \in [0,1]$ and $p \in [0,1]$ respectively denote the true and estimated activation probabilities.

When pre-training the classifier, it is only applied to the mixtures. In this case, the classification loss at frame τ is the weighted sum of binary cross-entropy terms over all sources,

$$\mathcal{L}_{\text{class}}(X,\tau) = \sum_{i=1}^{n} W_{i,\tau} H(l_{i,\tau}, p_{i,\tau}(X)), \tag{3}$$

where $l_{i,\tau}$, $p_{i,\tau}(.)$, and $W_{i,\tau}$ respectively denote the true label, estimated class probability, and the loss weight for the i-th source at time frame τ. The reason for weighting the loss terms is that there may be sound classes with very different activity levels in a dataset. For instance, a dataset of urban sounds might include rare sound events (e.g., gun shots) as well as sounds that are active over long periods of time (e.g., street music). In such scenarios, the weights should balance the class contributions to the total loss. We define the weights as

$$W_{i,\tau} = \begin{cases} \gamma_i^{-1} & \text{for } i \in \mathcal{A}_\tau, \\ (1-\gamma_i)^{-1} & \text{for } i \notin \mathcal{A}_\tau, \end{cases} \tag{4}$$

where γ_i is the activation probability for the i-th source and $\mathcal{A}_\tau$ the set of active source indices at time frame τ. We compute γ_i from training data as the ratio of the total number of frames where the i-th source is active to the total number of frames in the dataset. In the case with clip-level labels, a max-pooling operation is applied to the output of the frame classifier to map frame labels to clip labels. In this scenario, we do not have access to the frame-level prior knowledge regarding sound class activities, which hinders the use of class-related weights. The classification loss given the clip-level labels is thus formulated as

$$\mathcal{L}_{\text{class}}(X) = \sum_{i=1}^{n} H(l_i, p_i(X)) \tag{5}$$

where l_i and $p_i(.)$ are the clip-level true label and estimated class probability for the i-th source.

In training the separator, the classifier is applied to the source estimates (not to the mixture). The frame-level classification loss for each source estimate includes the associated true class label and zeros as true labels for all other sources:

$$\mathcal{L}_{\text{class}}(\hat{S}_i,\tau) = W_{i,\tau} H(l_{i,\tau}, p_{i,\tau}(\hat{S}_i)) + \sum_{j\neq i} W_{j,\tau} H(0, p_{j,\tau}(\hat{S}_i)), \tag{6}$$

which takes the following form for the clip-level case:

$$\mathcal{L}_{\text{class}}(\hat{S}_i) = H(l_i, p_i(\hat{S}_i)) + \sum_{j\neq i} H(0, p_j(\hat{S}_i))). \tag{7}$$

The overall loss when training the separator is obtained by combining the mixture magnitude and classification losses. The combined loss for the the frame-level labels is computed as

$$\mathcal{L}_{\text{total}} = \sum_{i,\tau} \mathcal{L}_{\text{class}}(\hat{S}_i,\tau) + \alpha \sum_{\tau} \mathcal{L}_{\text{mix}}(\tau), \tag{8}$$

and for the clip-level labels as

$$\mathcal{L}_{\text{total}} = \sum_{i} \mathcal{L}_{\text{class}}(\hat{S}_i) + \alpha \sum_{\tau} \mathcal{L}_{\text{mix}}(\tau), \tag{9}$$

where $\alpha \geq 0$ is a tunable parameter controlling the contribution of the mixture magnitude term to the total loss.

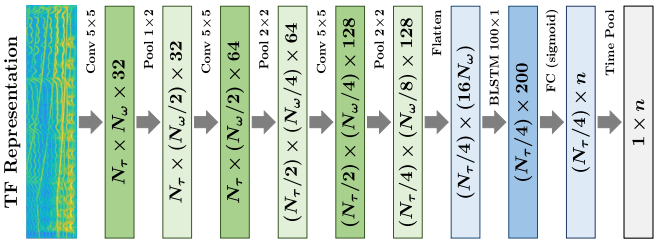

Fig. 2. Architecture of the 2D-CRNN classifier. N_τ and N_ω respectively denote the number of time frames and frequency bins in the input representation. n is the total number of sound classes.

2.3. Network Architecture

The separator block in our model consists of a 3-layer bidirectional long short-term memory (BLSTM) network, with each layer including 600 nodes in each direction. A fully connected layer maps the output of the BLSTM network to n masks with the same size as the input mixture. Activation functions of all BLSTM units are *tanh*, while the fully connected layer outputs go through *sigmoid* functions, so that the mask values are always in [0,1].

To design a frame-level classifier, we explored a number of architectures, ranging from very simple, such as a small stack of fully connected layers, to increasingly more sophisticated ones, such as convolutional recurrent neural networks (CRNNs). The clip-level classifier in this work is a simple extension of the frame-level classifier. It is built by adding a max-pooling operator to the output of the frame-level classifier for each sound class to perform frame-level to clip-level mapping of class probabilities. We leave the investigation of separation performance for some of the more advanced temporal pooling operations explored in [16] and [17] to future work.

Here, we present the architecture that performed best in our experiments: A CRNN architecture composed of a 3-layer 2D convolutional network including max-pooling after each layer, followed by a BLSTM layer, and a fully connected layer which maps the BLSTM output to class probabilities. Activation functions of convolutional, BLSTM, and fully connected layers are *relu*, *tanh*, and *sigmoid*, respectively. The output of each convolutional layer is batch normalized prior to the application of the activation function. Figure 2 illustrates this architecture in detail. This network is a slightly modified version of the SED model proposed in [17]. Note that the second and third pooling operations in the convolutional network are applied across both frequency and time axes, which results in a downsampled version of frame-level predicted probabilities. To match this coarser time resolution while computing the frame-level loss values, we also downsample the true weak labels via max-pooling.

3. EXPERIMENTS

In this section, we present the results of our experiments and compare our method to the common approach using strong labels.

3.1. Dataset

UrbanSound8K[1] [28] is a dataset of 8732 sound excerpts of length ≤ 4 s, taken from field recordings. The excerpts are labeled based on the sound event types and their salience in the auditory scene (foreground or background). The dataset contains 10 sound classes, from which we selected 5: car horn, dog bark, gun shot, jackhammer, and

[1] https://urbansounddataset.weebly.com/urbansound8k.html

Table 1. Mean SDR values (dB) ± standard deviation for all sound classes and separators trained using various labels. ΔSDR indicates the SDR improvement. The last column shows the results averaged over all samples and all classes.

	Sound class					
	Car horn	Dog bark	Gun shot	Jackhammer	Siren	Overall
Input SDR	-5.8 ± 5.1	-5.4 ± 4.8	-5.5 ± 4.4	-2.9 ± 4.8	-3.0 ± 4.6	-4.5 ± 4.9
ΔSDR-strong	9.9 ± 10.1	10.0 ± 7.1	12.5 ± 8.0	7.8 ± 6.6	4.9 ± 8.9	9.0 ± 8.6
ΔSDR-frame	7.0 ± 7.4	8.3 ± 5.6	9.7 ± 5.4	5.7 ± 4.2	3.1 ± 6.4	6.8 ± 6.3
ΔSDR-clip	6.5 ± 6.1	6.4 ± 4.4	8.8 ± 5.5	4.6 ± 3.8	1.8 ± 6.7	5.6 ± 5.9

siren. The class selection was made based on two criteria: i) examples in one class should contain mostly the sound corresponding to the class label, with a reasonable salience level, and ii) examples from different classes should be acoustically distinct enough so that they are at least recognizable as different sounds by humans.

Audio mixtures in our dataset are 4 seconds long and sampled at 16 kHz. Each mixture includes at least one *sound event* (i.e., a single occurrence of a sound class) from one of the five selected classes. Sound events are of arbitrary lengths, ranging from 0.5 s to 4 s, with a random (uniform) start time, all sound classes are sampled uniformly, and the level of each event ranges between -30 to -25 LUFS [29]. The total number of events per mixture is sampled from a zero-truncated Poisson distribution with an expected value of λ. Note that this number can include multiple sound events from one class, which are grouped together as one source while generating the weak labels from metadata. For instance, $\lambda = 5$, the value used in all our experiments, means that there are on average 5 sound events (from any class) per mixture. Our training, validation, and testing sets include 20K, 5K, and 5K mixtures, respectively.

3.2. Training Setup

The input to the separator is the log-magnitude STFT of a mixture. The input to the classifier is the linear magnitude STFT of a mixture or a source estimate. The features are generated using the square root of a *hann* window of length 32 ms and a hop size of 8 ms. The STFT parameters are the same for the separator and classifier inputs.

To provide an upper bound for the separation performance, we trained the separator network on strong labels (i.e., target sources). The joint separation-classification model was trained on either frame-level or clip-level weak labels, using $\alpha = 100$ based on a grid search for the best results. In all training sessions, we used the ADAM optimizer, with a learning rate of 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The batch size was set to 10 in all experiments. We train the networks until the loss on the validation set stops improving for 5 consecutive epochs, with a maximum of 50 epochs.

3.3. Results

We evaluate the performance of the classifier in terms of F-measure $\mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}}$, the harmonic mean of precision $\mathcal{P} = \frac{TP}{TP+FP}$ and recall $= \frac{TP}{TP+FN}$, where TP, FP, and FN respectively denote the number of true positives, false positives, and false negatives in classification results. Table 2 presents the average F-measure for frame-level and clip-level sound classification when the classifier is trained and tested on the mixtures.

To measure the quality of separated sources, we use the scale-invariant source to distortion ratio (SI-SDR) [30], which has been shown to be more appropriate for single-channel instantaneous separation evaluation than the original SDR [31]. Figure 3 shows the distribution of input and output SDRs for all classes and label types. The least and most amount of overlap between input and output SDR values are observed for the gun shot and siren classes, respectively. The siren class in our dataset contains a more diverse set of sounds compared to other classes (e.g., police siren versus ambulance siren), which is likely the reason it is the most difficult sound type to separate even when strong labels are used (see also Table 1). Although frame-level labels yield better results than clip-level labels in general, the distribution of output SDRs for these two label types appear highly overlapped in all cases. Furthermore, both weak-label distributions overlap reasonably well with the strong-label distributions and provide significant SDR improvement over the noisy mixtures.

Table 2. Sound source classification results in terms of F-measure.

	Sound class				
Label	Car horn	Dog bark	Gun shot	Jackhammer	Siren
Frame-level	0.948	0.870	0.856	0.940	0.876
Clip-level	0.958	0.924	0.949	0.922	0.914

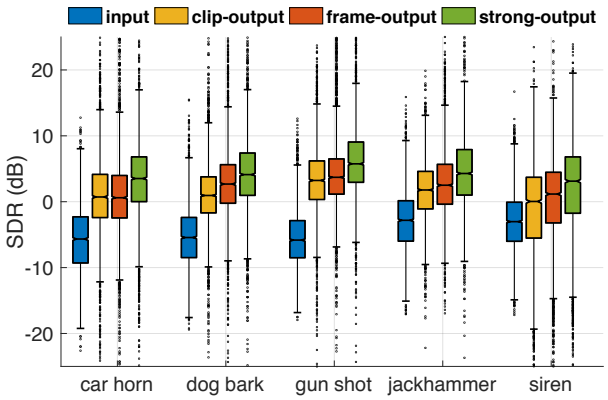

Fig. 3. Distribution of separation results for all sound classes. All boxes cover the values in the range of the first and third quartiles, with the middle notch indicating the median. For each source, box plots from left to right respectively correspond to the input SDR, and output SDRs using clip-level, frame-level, and strong labels.

4. CONCLUSION

We presented an approach to audio source separation using weak sound class labels. In our proposed model, an SED classifier is employed as the principal metric for loss calculation while training the separator. The model is trained to minimize an objective function that requires the classifier to identify the sound sources in the mixture as well as their isolated versions estimated by the separator. The objective function also enforces the estimated sources to sum to the mixture. Our experiments yielded promising results and showed significant SDR improvement even when using weak labels on a very coarse-resolution time grid.

5. REFERENCES

[1] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE ICASSP*, Apr. 2015.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, 2018.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, Mar. 2016.

[4] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, 2017.

[5] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE ICASSP*, Mar. 2017.

[6] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. IEEE IWAENC*, Sep. 2018.

[7] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some slakh:a dataset to study the impact of training data quality and quantity," in *Proc. IEEE WASPAA*, Oct. 2019.

[8] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A reference implementation for music source separation," *Journal of Open Source Software*, 2019.

[9] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE WASPAA*, Oct. 2019.

[10] E. Humphrey, S. Durand, and B. McFee, "OpenMIC-2018: An open data-set for multiple instrument recognition." in *Proc. ISMIR*, Sep. 2018.

[11] M. Cartwright, G. Dove, A. E. Méndez Méndez, J. P. Bello, and O. Nov, "Crowdsourcing multi-label audio annotation tasks with citizen scientists," in *Proc. ACM CHI*, 2019.

[12] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proc. ACM Multimedia*, Oct. 2016.

[13] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *Proc. IEEE ICASSP*, Mar. 2017.

[14] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. IEEE ICASSP*, Apr. 2018.

[15] 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 Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, 2019.

[16] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, 2018.

[17] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *Proc. IEEE ICASSP*, May 2019.

[18] B. Kim and B. Pardo, "Sound event detection using point-labeled data," in *Proc. IEEE WASPAA*, Oct. 2019.

[19] N. Zhang, J. Yan, and Y. Zhou, "Weakly supervised audio source separation via spectrum energy preserved Wasserstein learning," in *Proc. IJCAI*, Jul. 2019.

[20] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *Proc. IEEE ICASSP*, Apr. 2018.

[21] D. Stowell and R. E. Turner, "Denoising without access to clean data using a partitioned autoencoder," *arXiv preprint arXiv:1509.05982*, 2015.

[22] M. Michelashvili, S. Benaim, and L. Wolf, "Semi-supervised monaural singing voice separation with a masking network trained on synthetic mixtures," in *Proc. IEEE ICASSP*, May 2019.

[23] S. Ewert and M. B. Sandler, "Structured dropout for weak label and multi-instance learning and its application to score-informed source separation," in *Proc. IEEE ICASSP*, Mar. 2017.

[24] E. Karamatli, A. T. Cemgil, and S. Kirbiz, "Audio source separation using variational autoencoders and weak class supervision," *IEEE Signal Process. Lett.*, vol. 26, no. 9, 2019.

[25] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proc. ICCV*, Oct. 2019.

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

[27] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 4, 2019.

[28] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Multimedia*, Nov. 2014.

[29] E. Grimm, R. Van Everdingen, and M. Schöpping, "Toward a recommendation for a European standard of peak and LKFS loudness levels," *SMPTE Motion Imaging Journal*, vol. 119, no. 3, 2010.

[30] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. IEEE ICASSP*, May 2019.

[31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, 2006.