WEAKLY SUPERVISED SOURCE-SPECIFIC SOUND LEVEL ESTIMATION IN NOISY SOUNDSCAPES

Aurora Cramer, [1] *Mark Cartwright,*[2] *Fatemeh Pishdadian,*[3] *Juan Pablo Bello*[1]

[1] New York University, Music and Audio Research Laboratory, New York, NY, USA
[2] New Jersey Institute of Technology, Newark, NJ, USA
[3] Northwestern University, Evanston, IL, USA

ABSTRACT

While the estimation of what sound sources are, when they occur, and from where they originate has been well-studied, the estimation of how loud these sound sources are has been often overlooked. Current solutions to this task, which we refer to as source-specific sound level estimation (SSSLE), suffer from challenges due to the impracticality of acquiring realistic data and a lack of robustness to realistic recording conditions. Recently proposed weakly supervised source separation offer a means of leveraging clip-level source annotations to train source separation models, which we augment with modified loss functions to bridge the gap between source separation and SSSLE and to address the presence of background. We show that our approach improves SSSLE performance compared to baseline source separation models and provide an ablation analysis to explore our method's design choices, showing that SSSLE in practical recording and annotation scenarios is possible.

Index Terms— machine listening, source separation, sound event recognition, source-specific sound level estimation, weakly supervised learning

1. INTRODUCTION

The field of machine listening is concerned with addressing the perception and characterization of acoustic scenes and their constituent sound events. Sound event recognition (SER) is one of the fundamental problems of machine listening, which is concerned with *what* sound sources are present in a soundscape and when they occur [1]. Sound source localization is another fundamental problem concerned with from *where* sound sources originate [2, 3]. Both of these problems are well studied. Sound level estimation for specific sources, referred to henceforth as *source-specific sound level estimation (SSSLE)*, is concerned with *how loud* sound sources are, but is not well-studied in the case of sources in polyphonic soundscapes.

Despite being under-studied, SSSLE has many important yet often unrealized applications. In urban noise pollution monitoring, SSSLE could estimate the loudness of specific sound sources (e.g., traffic or construction equipment) to aid in noise mapping and enforcement [4, 5, 6, 7]. In intelligent audio production, SSSLE is used to determine the relative gain of instruments in audio mixes and inform automatic mixing systems that mimic audio engineers [8, 9]. SSSLE could also be used to aid in distance estimation for diverse source localization applications such as wildlife monitoring [10] and sound awareness technology for the hearing impaired [11].

Please direct correspondence to jtc440@nyu.edu
This work is partially supported by National Science Foundation award 1633259 (BIRDVOX) and award 1544753 (SONYC).

Existing SSSLE approaches suffer from the challenge of collecting reliable ground truth for developing and evaluating methods. Measuring source-specific sound levels for ground truth data is nontrivial in realistic (e.g., noisy and polyphonic) settings such as city streets. While source separation techniques can help to isolate sources in mixtures, these techniques do not perform well in background noise and out-of-vocabulary sources nor formulated to directly address SSSLE.

In this work, we propose a framework for SSSLE that not only can be trained with just clip-level source presence annotations but also *bridges the gap between source separation and SSSLE* and *accounts for background*, both of which we posit are important steps to advance SSSLE. We evaluate our proposed models on a synthetic dataset of urban sound sources with realistic noise and out-of-vocabulary sources. We find that they outperform baseline source separation methods, and an ablation analysis investigates the effect of design choices in our framework. While our approach has been developed and evaluated using a synthetic dataset, one of its strengths is that it can be applied to real, noisy recordings with just clip-level annotations.

2. RELATED WORK

2.1. Sound level estimation

Typically, sound levels are measured for an entire audio signal, regardless of the sources, and can be measured at short-term scales or in an integrated fashion considering the entire signal. For digital audio signals, full-scale sound level measures such as *dBFS (decibels relative to full scale)* [12] are typically used and can be calibrated to estimate sound pressure levels. In contrast, loudness measures like *LUFS (loudness units relative to full scale)* apply perceptual weighting as well as silence gating mechanisms to provide sound level estimates more closely related to perceived loudness [13].

Prior research in SSSLE has primarily focused on estimating traffic sound levels in noise monitoring and the estimation of isolated instrument track levels in automatic mixing systems. Traditional approaches to SSSLE require access to isolated sources [8, 14, 15]. However, more recent approaches have used fully-supervised source separation techniques to perform well in the presence of other known, modeled sources [5, 6, 7, 9].

Using source separation [16] to isolate a specific source and directly measure its sound levels is a straightforward approach to SSSLE. If we have a perfect source separation system (assuming no scale normalization), then we can perfectly estimate the sound level of each source. Time-frequency masking based deep learning methods are among the most popular source separation approaches,

particularly because of the flexibility afforded by modern deep learning frameworks. However, while these methods perform fairly well for source separation, they suffer the drawback of requiring time-frequency-level ground-truth obtained from isolated recordings. Since obtaining isolated recordings in realistic soundscapes is generally impractical or impossible, training these models on realistic soundscapes can prove difficult [17]. Fortunately, recent methods have addressed this impracticality by training models using weaker supervision, via clip-level or frame-level annotations [18]. We therefore restrict the scope of this paper to these methods.

2.2. Discriminator-based weakly supervised source separation

Discriminator-based weakly supervised methods for source separation train a model to separate sources in a mixture using a classifier to critique the separated sources [18, 19, 20] using only clip-level or frame-level annotations as well as *energy consistency* between the mixture and the reconstructed sources [18]. More recent methods leverage unsupervised methods involving learning to mix and separate mixtures [21, 22], though we restrict the scope of this work to weakly supervised methods. We build upon on the framework presented by Pishdadian et al. [18], focusing on the clip-level scenario since it reflects practical SSSLE settings, and describe how *energy consistency loss* and *classification loss* enable weak supervision.

Formally, let $\mathbf{X} \in \mathbb{R}_{\geq 0}^{F \times T}$ be the time-frequency magnitude representation for a sound mixture and let $\mathbf{y} = [y_1, \ldots, y_C]^\intercal \in \{0,1\}^C$ be the clip-level class presence vector. In this framework, a source separation network parameterized by θ takes $\mathbf{X}$ as input and produces a time-frequency mask $\hat{\mathbf{M}}_i = \mathbf{f}_{i,\theta}(\mathbf{X}) \in [0,1]^{F \times T}$ for class i, resulting in the source estimate $\hat{\mathbf{S}}_i = \hat{\mathbf{M}}_i \odot \mathbf{X}$.

An *energy consistency* loss $\mathcal{L}_{\text{mix}}$ is applied to ensure that estimated *active sources* (classes with $y_i = 1$) contain the same energy as the mixture, and that estimated *inactive sources* (classes with $y_i = 0$) contain no energy. Let $\mathcal{A}_+$ be the ground-truth set of active sources in the clip. We define the energy consistency residual for active sources, $\mathbf{R}_{\text{active}} = \mathbf{X} - \sum_{i \in \mathcal{A}_+} \hat{\mathbf{S}}_i$, and the energy consistency residual for inactive sources, $\mathbf{R}_{\text{inactive}} = \sum_{j \notin \mathcal{A}_+} \hat{\mathbf{S}}_j$. Then we have:

$$\mathcal{L}_{\text{mix}} = \mathcal{L}_{\text{mix-active}} + \mathcal{L}_{\text{mix-inactive}} \tag{1}$$

$$= \frac{1}{TF} \left\| \mathbf{M}_E \odot \mathbf{R}_{\text{active}} \right\|_1 + \frac{1}{TF} \left\| \mathbf{M}_E \odot \mathbf{R}_{\text{inactive}} \right\|_1 \tag{2}$$

where $\mathbf{M}_E \in \{0,1\}^{F \times T}$ is a masking matrix that zeros out frames containing less than 1% of the maximum frame energy in the clip. By applying this mask, we only train on audio containing salient activity; whenever we normalize by T, we actually normalize by the number of salient frames, but omit this distinction for brevity. Additionally, we use $\| \cdot \|_1$ to indicate the element-wise ℓ_1 norm; that is, $\|\mathbf{X}\|_1 = \sum_{i,j} |X_{ij}|$.

A classifier network then independently takes in each estimated $\hat{\mathbf{S}}_i$ as input, which should detect only the presence of class i when indeed present. We also expect that when given the mixture $\mathbf{X}$ as input, the classifier's estimates for each class should match the ground-truth. Specifically, we apply the *classification loss*:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{cls-mix}} + \sum_i \mathcal{L}_{\text{cls-mix},i} \tag{3}$$

where $\mathcal{L}_{\text{cls-mix}}$ indicates binary cross-entropy between clip-level mixture targets y_i and mixture classifier predictions $\hat{y}_i^{(\text{mix})}$ and $\mathcal{L}_{\text{cls-mix},i}$ indicates binary cross-entropy between a modified target $\mathbf{y} \odot \mathbf{e}_i$ (where $\mathbf{e}_i$ is a canonical basis vector) and the clip-level classifier prediction for the separated source for class i. Note that when training the separator, $\mathcal{L}_{\text{cls-mix}}$ only has an effect if also optimizing the classifier. However, there is evidence that independently training the classifier and fixing the weights prior to training the separator results in improved separation [18]; therefore, we follow this procedure. The final loss for this framework is then:

$$\mathcal{L}_{\text{weak}} = \alpha \mathcal{L}_{\text{mix}} + \mathcal{L}_{\text{cls}} \tag{4}$$

where α is a hyperparameter determining the relative importance of each term. While this framework affords us the ability to train source separation models with more easily attainable annotations, which can make SSSLE more practical, it does not directly address SSSLE nor does it account for the presence of background noise or out-of-vocabulary events. Therefore, we propose an augmented framework to address these fundamental concerns.

3. SOURCE-SPECIFIC SOUND LEVEL ESTIMATION

3.1. Bridging source separation and SSSLE

The energy consistency terms in the loss for weakly supervised source separation are in the form $\frac{1}{TF}\|\mathbf{R}\|_1$. However, this term enforces consistency at the time-frequency scale, whereas we are ultimately interested in estimating the sound level. Note that the sound level can be estimated from either the energy spectrum or from the total energy of the signal. However, the classifier requires a time-frequency input to properly critique the separator estimates. Therefore, we can augment the energy consistency terms to enforce energy consistency at the spectrum level or the global energy level using the time-frequency error terms. For convenience, we call such augmentations *sound level augmentations*. We propose an augmentation to this loss of the form:

$$\frac{1}{TF_{\text{fb}}} \|h_\Phi(\mathbf{R})\|_1 = \frac{1}{TF_{\text{fb}}} \|\mathbf{B}_L \mathbf{A} \mathbf{R} \mathbf{B}_R\|_1 \tag{5}$$

where $\mathbf{A} \in \mathbb{R}_{\geq 0}^{F_{\text{fb}} \times F}$ is a filter bank matrix with F_{fb} frequency bands, $\mathbf{B}_L \in \mathbb{R}_{\geq 0}^{F_{\text{out}} \times F_{\text{fb}}}$ is a frequency aggregation matrix producing F_{out} frequency bands, $\mathbf{B}_R \in \mathbb{R}_{\geq 0}^{T \times T_{\text{out}}}$ is a temporal aggregation matrix producing T_{out} frames, and $\Phi = (\mathbf{A}, \mathbf{B}_L, \mathbf{B}_R)$. This parameterization allows energy error to be aggregated to enforce energy consistency at different time-frequency resolutions. $\mathbf{A}$ is used to implement a filter bank transformation (in the frequency domain) that can be applied to emphasize reconstruction in perceptually relevant frequency bands. In particular, we consider a mel frequency filter bank $\mathbf{A}_{\text{mel}}$. If no filter bank is used, $\mathbf{A} = \mathbf{A}_{\text{linear}} = \mathbf{I}_F$. We consider the following key choices of $\mathbf{B}_L$ and $\mathbf{B}_R$:

- *time-frequency energy consistency*: $\mathbf{B}_L = \mathbf{I}_F, \mathbf{B}_R = \mathbf{I}_T$
- *spectrum energy consistency*: $\mathbf{B}_L = \mathbf{I}_F, \mathbf{B}_R = \mathbf{1}_T$
- *global energy consistency*: $\mathbf{B}_L = \mathbf{1}_F^\intercal, \mathbf{B}_R = \mathbf{1}_T$

While using only the global energy consistency configuration or even spectrum energy consistency configuration would most closely match the end goal of sound level estimation, we would lose the time-frequency structure afforded by the time-frequency energy consistency configuration. The classifier still enforces time-frequency structure by classifying the estimated spectrograms; however, without the time-frequency consistency loss there is less incentive for the estimated sources to resemble the original mixture. Therefore,

we propose enforcing mean energy consistency across multiple time-frequency resolutions. Specifically, for a set of parameterizations $\mathcal{P} = \{\Phi_1, \ldots, \Phi_P\}$, we can take the mean energy consistency:

$$\frac{1}{|\mathcal{P}|} \sum_{\Phi \in \mathcal{P}} \frac{1}{TF_{\text{fb}}} \|h_\Phi(\mathbf{R})\|_1 \tag{6}$$

With this method, we can easily enforce energy consistency at various resolutions to retain time-frequency structure while directly addressing sound level estimation. We therefore choose $\mathcal{P}_{\text{all, mel}} = \{\Phi_{\text{time-freq, mel}}, \Phi_{\text{spectrum, mel}}, \Phi_{\text{global}}\}$ to incorporate multi-resolution structure and choose $\mathbf{A}_{\text{mel}}$ focus on perceptually relevant frequency bands with mel frequency filter banks. However, for global energy consistency we always use $\mathbf{A}_{\text{linear}}$ to avoid redundancy.

3.2. Accounting for background

The energy consistency term must be modified in the presence of background noise and out-of-vocabulary events, since the sum of the sources of interest will not in general result in the mixture. For convenience, we call such augmentations *background augmentations*. A straightforward solution is to allow for the sum of source energy to underestimate the mixture energy. To achieve this, we can again modify Eq. 5 using an asymmetric margin parameterized by $\varepsilon \geq 0$:

$$\|\mathbf{R}\|_1^{(\text{asym}, T, F, \varepsilon)} = \left[\left\| [\mathbf{R}]_+ \right\|_1 - TF\varepsilon \right]_+ + \left\| [-\mathbf{R}]_+ \right\|_1 \tag{7}$$

where $[\,\cdot\,]_+$ indicates element-wise half-wave rectification. Note that ε specifies the allowable mean energy per time-frequency bin and that when $\varepsilon = 0$, $\|\mathbf{R}\|_1^{(\text{asym}, \varepsilon)} = \|\mathbf{R}\|_1$. A reasonable choice for ε is the mean energy margin estimated from the training set. We then have a residual signal containing only background and out-of-vocabulary sources, and therefore the classifier should indicate that no in-vocabulary sources are present. To the residual spectrogram estimate $\hat{\mathbf{S}}_{\text{bkgr}} = \left[1 - \sum_i \hat{\mathbf{M}}_i\right]_+ \odot \mathbf{X}$, we apply the loss:

$$\mathcal{L}_{\text{cls-bkgr}} = \sum_i H\left(0, \hat{y}_i^{(\text{bkgr})}\right) \tag{8}$$

where $\hat{y}_i^{(\text{bkgr})}$ is the clip-level classifier prediction for class i for the residual spectrogram $\hat{\mathbf{S}}_{\text{bkgr}}$ and H is the binary cross-entropy function. By allowing for a residual signal and by enforcing that it does not contain any in-vocabulary sound sources, we can handle additive background noise and out-of-vocabulary sources in a principled way.

3.3. Putting it all together

Combining the aforementioned augmentations with the approach detailed in Section 2.2, we have the components to build our framework for SSSLE. Our SSSLE model optimizes the loss:

$$\mathcal{L}_{\text{weak, sssle}}^{\mathcal{P}} = \frac{\alpha}{|\mathcal{P}|} \sum_{\Phi \in \mathcal{P}} \mathcal{L}_{\text{mix, sssle}}^{\Phi} + \mathcal{L}_{\text{cls, sssle}} \tag{9}$$

where we have

$$\mathcal{L}_{\text{mix, sssle}}^{\Phi} = \mathcal{L}_{\text{mix-active, sssle}}^{\Phi} + \mathcal{L}_{\text{mix-inactive, sssle}}^{\Phi} \tag{10}$$

$$\mathcal{L}_{\text{mix-active, sssle}}^{\Phi} = \frac{1}{TF_{\text{fb}}} \|h_\Phi(\mathbf{M}_E \odot \mathbf{R}_{\text{active}})\|_1^{(\text{asym}, T, F_{\text{fb}}, \varepsilon)} \tag{11}$$

$$\mathcal{L}_{\text{mix-inactive, sssle}}^{\Phi} = \frac{1}{TF_{\text{fb}}} \|h_\Phi(\mathbf{M}_E \odot \mathbf{R}_{\text{inactive}})\|_1 \tag{12}$$

$$\mathcal{L}_{\text{cls, sssle}} = \mathcal{L}_{\text{cls-mix}} + \sum_i \mathcal{L}_{\text{cls-mix}, i} + \beta \mathcal{L}_{\text{cls-bkgr}} \tag{13}$$

where $\beta \in \{0, 1\}$. When $\beta = 1$, $\mathcal{L}_{\text{cls-bkgr}}$ is included in the loss, which ensures that the residual signal (containing the background) does not contain the presence of any of the in-vocabulary classes. With this framework, we can more directly address SSSLE within the weakly supervised source separation framework.

4. EXPERIMENTAL METHODS

The procedure we use to evaluate our proposed methods is as follows. First, we generate a dataset of soundscapes containing sound events and varying levels of background noise, with accompanying clip-level source annotations. We then train models with our proposed loss augmentations, compare them to baseline source separation models, and perform an ablation analysis on our method's design choices. To evaluate source separation performance, we use scale-invariant signal-to-noise ratio (SI-SDR) [23] improvement with respect to the baseline of using the mixture as the estimate. The reconstructed audio signals are obtained by taking the ISTFT of the masked spectrogram with the mixture phase. To evaluate sound level estimation, we use the absolute error with respect to dBFS. In the proceeding sections, we describe the components of our experiments.

4.1. Dataset creation

To train and evaluate these models, we use the synthetic dataset of urban soundscapes[1] used by Pishdadian et al. [18]. This dataset contains 4 second synthetic mixtures sampled at 16 kHz containing foreground events from UrbanSound8K [24], specifically from the 5 classes *car horn*, *dog bark*, *gun shot*, *jackhammer*, and *siren*. Folds 1–6, 7–8, and 9–10 of UrbanSound8K are used to generate the training, validation, and testing subsets, respectively, with 50k, 10k, and 10k examples each. Each soundscape contains a number of events sampled from a zero-truncated Poisson distribution ($\lambda = 5$), with a uniformly random start-time and a uniformly chosen class.

To evaluate the methods in the presence of background noise and potentially out-of-vocabulary events, we add background noise to the base dataset at varying sound levels. For the background audio, we use audio clips recorded by the Sounds of New York City acoustic sensor network [25] that an urban sound tagging classifier identified as not containing urban sound classes. The classifier was trained on SONYC-UST-V1 [26] and uses OpenL3 embeddings [27] as input. We have released these background clips in the SONYC-Backgrounds[2] dataset. We create a train/valid/test split by choosing disjoint sensors resulting in a roughly 60/20/20 split. For each example in the base dataset, we uniformly sample a 4-second segment of a background clip, which is subsequently LUFS-normalized mixed with the original example. We create datasets using background sound levels $\in \{-50, -20, 0\}$ dB LUFS, which we refer to as *weak* background, *moderate* background, and *strong* background. We have made these datasets available for the sake of reproducibility.[3]

4.2. Baseline comparisons

For our baseline, we use the models without the augmentations outlined in Section 2.2. We also compare with models individual and combined effects of sound level augmentations and background augmentations. Our proposed models are trained to minimize $\mathcal{L}_{\text{weak, sssle}}^{\mathcal{P}}$,

[1]Thanks to Gordon Wichern at Mitsubishi Electric Research Laboratory for their correspondence and use of their data.

[2]SONYC-Backgrounds: https://doi.org/10.5281/zenodo.5129078

[3]Soundscape data: https://doi.org/10.5281/zenodo.5123372

with $\mathcal{P} = \mathcal{P}_{\text{all, mel}}$ and $\beta = 1$. We choose $\alpha = 100$ to remain consistent with previous work [18]. Classifiers are trained separately to minimize $\mathcal{L}_{\text{cls-mix}}$, with classifier weights are frozen when training the separation model. ε is estimated using the empirical mean of time-frequency margins between active sources and mixtures from training examples. All separation models are trained with all levels of background noise (none, weak, moderate, and strong).

4.3. Ablation experiments

We perform two ablation studies to explore the design choices in the sound level and background augmentations. For the sound level augmentations, we look at all combinations of time-frequency, spectrum, and global energy consistency loss as well as the use of mel frequency bands. Since background noise presence is not central to these augmentations, we restrict our attention to models trained on noise-less mixtures. For the background augmentations, we disable residual background classification and remove the energy margin to explore their effect on model performance in background noise.

4.4. Training details

For the front-end to the models, we use a log-magnitude STFT with a DFT-size of 512 with 25% overlap with a square-root Hann window applied. When mel frequency filter banks are used, we use 40 bands. For the source separation models, instead of the BLSTM used by Pishdadian et al. [18], we use a variant of the popular UNet model used by Kong et al. [22], removing the conditioning mechanism and increasing the number of outputs to the number of sound sources. For the classification models, we use the linear-frequency CRNN model used by Pishdadian et al. [18], training it on the base dataset with no added background. All models are trained with the Adam optimizer, with initial learning rate 10^{-4} and batch size 8, for up to 50 epochs using early stopping on the validation set with a patience of 5 epochs. Our training code is available on our GitHub repository[4].

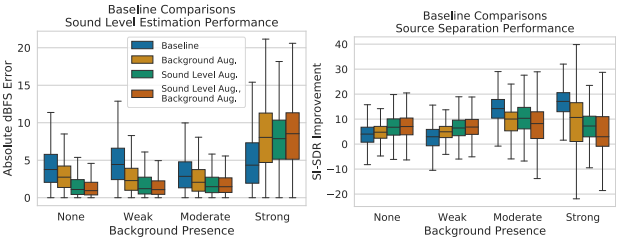
5. RESULTS AND DISCUSSION

5.1. Baseline comparison results

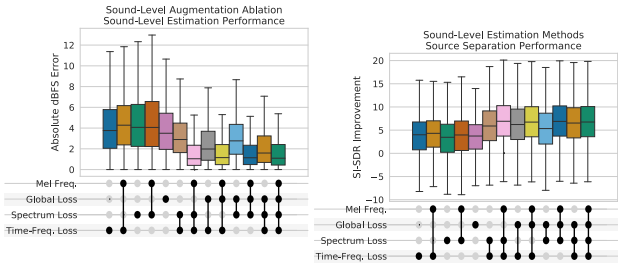
We see in Figure 1(a) that our proposed augmentations improve SSSLE performance and that using both sound-level and background augmentations provide the best performance. We see significant performance improvement in background noise, though our approach still fails in high noise scenarios. This may be because with a sufficiently large margin, the mixture loss no longer provides enough structure to reconstruct sources in the mixture. Source separation performance also improves in low noise conditions with our augmentations, though the baseline performs better with stronger noise, showing that our methods can improve source separation performance in some cases.

5.2. Ablation analysis

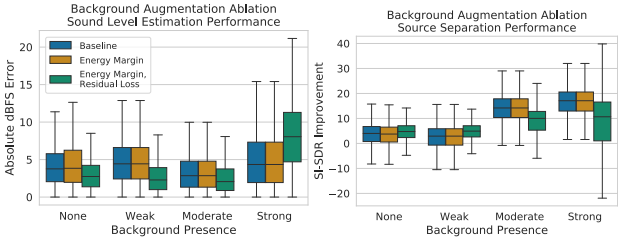
From the ablation analysis of the sound level augmentations shown in Figure 1(b), we see that using multiple time-frequency resolutions improved sound level estimation, though global energy consistency helps less than spectrum energy consistency. This may indicate that while temporal aggregation helps, constraining frequency structure is still important. Additionally, using mel frequency scales in all cases is more effective. From the ablation analysis of the background augmentations shown in Figure 1(c), introducing an energy margin and classifying the residual improve performance together.



(a) Baseline comparison results



(b) Sound level augmentation ablation study results



(c) Background augmentation ablation study results

Figure 1: Boxplots of evaluation metrics across test examples (outliers omitted) for models in each set of experiments. (left): Sound level estimation performance w.r.t. absolute dBFS error. Lower is better. (right): Source separation performance w.r.t. SI-SDR improvement. Higher is better.

6. DISCUSSION AND FUTURE PERSPECTIVES

While the results are preliminary, we have shown that extending weakly supervised source separation methods to directly address sound level estimation and to handle background noise improves the use of such systems for SSSLE, showing great promise for the use of clip-level annotations for SSSLE in realistic recording scenarios. While in this study we trained and evaluated using synthetic mixtures, real recordings of soundscapes with clip-level annotations could also be used. Evaluating SSSLE performance with a qualitative sound level metric like dBFS generally remains an open question, since obtaining ground truth sound level measurements is generally impossible in realistic scenarios. Subjective listening tests could be used to judge source-specific loudness, though this requires further development. By showing that SSSLE is possible in realistic recording scenarios with only clip-level annotations, we hope to engender enthusiasm and research around moving forward SSSLE.

[4]Code: https://github.com/sonyc-project/weakly-supervised-sssle

7. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, Eds., *Computational Analysis of Sound Scenes and Events*. Springer International Publishing, 2018.

[2] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[3] N. Madhu, R. Martin, U. Heute, and C. Antweiler, "Acoustic source localization with microphone arrays," *Advances in Digital Speech Transmission*, pp. 135–170, 2008.

[4] H. Di, X. Liu, J. Zhang, Z. Tong, M. Ji, F. Li, T. Feng, and Q. Ma, "Estimation of the quality of an urban acoustic environment based on traffic noise evaluation models," *Applied Acoustics*, vol. 141, pp. 115 – 124, 2018.

[5] J.-R. Gloaguen, A. Can, M. Lagrange, and J.-F. Petiot, "Road traffic sound level estimation from realistic urban sound mixtures by non-negative matrix factorization," *Applied Acoustics*, vol. 143, pp. 229 – 238, 2019.

[6] J.-R. GLOAGUEN, A. Can, J.-F. PETIOT, and M. Lagrange, "Study of the Non-negative Matrix Factorization behavior to estimate the urban traffic sound levels," in *ICSV'26 ; 26th International Congress on Sound and Vibration*, MONTREAL, Canada, July 2019, pp. –.

[7] J.-R. Gloaguen, A. Can, M. Lagrange, and J.-F. Petiot, "Estimating traffic noise levels using acoustic monitoring: a preliminary study," in *DCASE 2016, Detection and Classification of Acoustic Scenes and Events*, BUDAPEST, Hungary, Sept. 2016, p. 4p, dCASE 2016, Detection and Classification of Acoustic Scenes and Events.

[8] J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim, "Automatic multi-track mixing using linear dynamical systems," in *in Proceedings of the 8th Sound and Music Computing Conference*, 2011.

[9] D. Ward, H. Wierstorf, R. Mason, M. Plumbley, and C. Hummersone, "Estimating the loudness balance of musical mixtures using audio source separation," in *Proceedings of the 3rd Workshop on Intelligent Music Production (WIMP 2017)*, 2017.

[10] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PLOS ONE*, vol. 14, no. 10, p. e0214168, 2019.

[11] D. Jain, L. Findlater, J. Gilkeson, B. Holland, R. Duraiswami, D. Zotkin, C. Vogler, and J. E. Froehlich, "Head-mounted display visualizations to support sound awareness for the deaf and hard of hearing," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 241–250.

[12] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics, 4th Edition*, 1999.

[13] 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, pp. 28–34, 2010.

[14] C. Don and I. Rees, "Road traffic sound level distributions," *Journal of Sound and Vibration*, vol. 100, no. 1, pp. 41 – 53, 1985.

[15] E. Perez-Gonzalez and J. Reiss, "Automatic gain and fader control for live mixing," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 1–4.

[16] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, 1st ed. Wiley Publishing, 2018.

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

[18] F. Pishdadian, G. Wichern, and J. Le Roux, "Finding strength in weakness: Learning to separate sounds with weak supervision," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2386–2399, 2020.

[19] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint separation-classification model for sound event detection of weakly labelled data," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 321–325.

[20] 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 Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 777–787, 2019.

[21] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised speech separation using mixtures of mixtures," in *ICML 2020 Workshop on Self-supervision in Audio and Speech*, July 2020.

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

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

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

[25] J. P. Bello, C. Silva, O. Nov, R. L. DuBois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[26] M. Cartwright, A. E. M. Mendez, J. Cramer, V. Lostanlen, G. Dove, H.-H. Wu, J. Salamon, O. Nov, and J. Bello, "Sonyc urban sound tagging (sonyc-ust): a multilabel dataset from an urban acoustic sensor network," 2019.

[27] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.