

IS³ : Generic Impulsive–Stationary Sound Separation in Acoustic Scenes using Deep Filtering

Clémentine Berger¹, Paraskevas Stamadiatis¹, Roland Badeau¹, Slim Essid¹

¹LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Abstract—We are interested in audio systems capable of performing a differentiated processing of stationary backgrounds and isolated acoustic events within an acoustic scene, whether for applying specific processing methods to each part or for focusing solely on one while ignoring the other. Such systems have applications in real-world scenarios, including robust adaptive audio rendering systems (e.g., EQ or compression), plosive attenuation in voice mixing, noise suppression or reduction, robust acoustic event classification or even bioacoustics. To this end, we introduce IS³, a neural network designed for Impulsive–Stationary Sound Separation, that isolates impulsive acoustic events from the stationary background using a deep filtering approach, that can act as a pre-processing stage for the above-mentioned tasks. To ensure optimal training, we propose a sophisticated data generation pipeline that curates and adapts existing datasets for this task. We demonstrate that a learning-based approach, build on a relatively lightweight neural architecture and trained with well-designed and varied data, is successful in this previously unaddressed task, outperforming the Harmonic–Percussive Sound Separation masking method, adapted from music signal processing research, and wavelet filtering on objective separation metrics.

1. INTRODUCTION

An acoustic scene can be roughly decomposed into a stationary ambient background, containing a mixture of environmental sounds (wind, rain, insects etc.) and anthropogenic noises (traffic noise, speech babble noise or murmur, ventilation noise etc.), overlaid with isolated and *impulsive* acoustic events. These impulsive events are characterized by a sudden increase in sound pressure level over a short duration and can stand out to varying degrees from the background. Examples include impacts, explosions, bursts, clapping, short alarms, or even coughing... In many contexts, these two categories of sounds, stationary ambient backgrounds and impulsive events, require distinct and independent processing due to their differing characteristics. This is particularly relevant in audio mixing (e.g., differentiated equalization and compression) or audio pre-processing for tasks such as speech enhancement, and noise reduction/suppression.

Related works. To enable such a differentiated processing, separating the stationary background from impulsive sounds may be beneficial, allowing for targeted treatments. However, this specific separation task remains under-explored in the literature. Existing approaches primarily focus on impulsive noise attenuation or suppression for specific applications such as music restoration [1], [2], speech communication [3]–[7], and specialized domains such as automotive or aeronautical noise reduction [8], [9] and bioacoustics [10]. These methods often target context-specific noise (including audio artefacts rather than distinct sound events), and rely mostly on signal processing techniques that first detect impulsive events and subsequently remove them using interpolation and magnitude adjustment techniques [7] or separate them through reconstruction methods [11], masking [2], or wavelet filtering [1], [6], [8].

In contrast, we focus on general acoustic scenes from everyday life, aiming to separate and reconstruct both ambient backgrounds and

the impulsive sounds as faithfully as possible to support downstream applications. Blind Source Separation (BSS) methods appear well-suited for this task. Some studies have explored matrix demixing using statistical signal analysis [12], while others have focused on time-frequency (TF) domain masking approaches. Notably, in the musical domain, the Harmonic–Percussive Source Separation (HPSS) method [13] has been proposed, leveraging median filtering along both the time and frequency axes to generate harmonic and percussive masks for source separation.

More recently, deep learning approaches, particularly deep filtering, have surpassed traditional ratio-masking for speech enhancement [14], estimating complex-valued time-frequency filters that captures correlations with adjacent TF bins and improving the extraction process. However, this comes at the cost of increased computational complexity. To address this, DeepFilterNet [15], [16] balances deep filtering and real-gain predictions on an equivalent rectangular bandwidth (ERB) spectral representation, achieving state-of-the-art performance while remaining lightweight for real-time applications.

Contributions. We propose IS³, a deep filtering approach for Impulsive–Stationary Sound Separation in ambient acoustic scenes, aimed at reconstructing both impulsive components and the stationary background. A key challenge in this task is obtaining high-quality training data for supervised learning, which requires a diverse set of clean acoustic scenes free from impulsive sounds, combined with an extensive variety of isolated impulsive sounds. Our contributions are: i) a methodology for curating and adapting existing datasets to this task, along with a procedure for generating training, validation, and test data by combining these datasets; ii) a learning-based approach for the task of impulsive–stationary sound separation build on the adaptation of the DeepFilterNet architecture [16]; iii) an extensive evaluation on realistic data showing the superiority of our system to previous approaches including the HPSS masking method and an adaptation of the wavelet-based process from Nongpiur’s article [6].

2. MODEL

We first provide an overview of our system IS³, followed by a description of the loss functions used for optimization.

2.1. System overview

The architecture chosen for IS³ is strongly inspired by that of DeepFilterNet [15], [16] for speech enhancement, adapted here for impulsive–stationary sound separation as presented in Figure 1. The model follows an encoder-decoder structure that predicts parameters for a two-stage filtering process, corresponding to varying levels of filtering precision. The first stage predicts real-valued gains defined on ERB frequency bands, while the second stage performs deep filtering (DF) by predicting a complex filter.

The IS³ system takes as input an acoustic signal $x(t)$ sampled at 44100 Hz, which we decompose as follows:

$$x(t) = y_i(t) + y_s(t), \quad (1)$$

This work was supported by the Audible project funded by French BPI, and was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011014883R1).

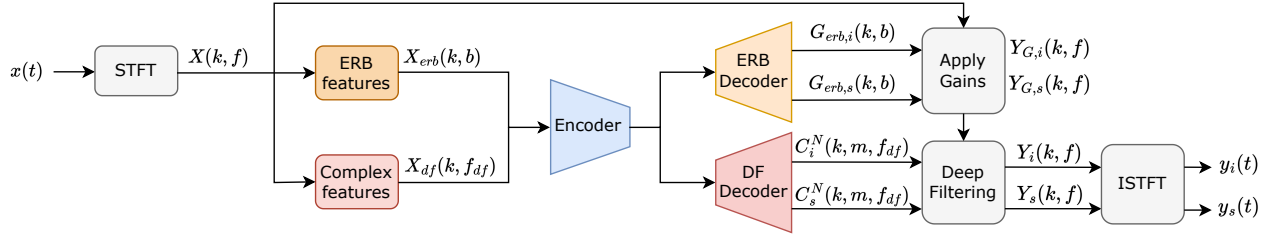


Fig. 1: Overview of the IS³ system: The model extracts real ERB and complex features from the input mixture and processes them through a shared encoder. Two decoders then operate in parallel: one predicts real-valued ERB gains for impulsive and stationary components, while the other estimates complex time-frequency filters for each source. These gains and filters are applied to the input spectrogram, and the separated signals are reconstructed via inverse STFT.

where y_s represents the stationary part of the acoustic scene and y_i denotes the impulsive components. The separation process operates in the frequency domain:

$$X(k, f) = Y_i(k, f) + Y_s(k, f), \quad (2)$$

where X is the short-time fourier transform (STFT) of x , computed with an analysis window of $N_{fft} = 2048$ with a 75% overlap, and a Hanning window. The indices k and f represent time and frequency bins, respectively.

The encoder processes both magnitude and complex features, as described in [15]. Magnitude features, denoted by $X_{erb}(k, b)$, $b \in [1, N_{erb}]$, are extracted using a rectangular ERB filterbank with N_{erb} bands applied to the normalized log-power spectrogram. Complex features, $X_{df}(k, f')$, are obtained by extracting the first N_{feat} frequency bands up to the frequency f_{df} from the complex spectrogram and applying band-wise unit normalization.

An ERB decoder then converts this information into predicted gains for each ERB band: $G_{erb,i}(k, b)$ and $G_{erb,s}(k, b)$, corresponding to the impulsive and stationary background components, respectively. An inverse ERB filterbank is applied to these gains, which are then used to filter the input spectrogram, yielding partially extracted impulsive $Y_{G,i}(k, f)$ and stationary $Y_{G,s}(k, f)$ spectrograms. This first filtering stage provides an initial coarse processing, which is then refined by the deep filtering step.

A DF decoder predicts two complex filters, C_i^M and C_s^M , applied up to the N_{feat} -th frequency band of the spectrograms obtained after the ERB stage. These filters separate the impulsive and stationary spectrograms $\hat{Y}_i(k, f)$ and $\hat{Y}_s(k, f)$:

$$\hat{Y}_i(k, f') = \sum_{m=0}^M C_i^M(k, m, f') \cdot Y_{G,i}(k - m, f'), \quad (3)$$

$$\hat{Y}_s(k, f') = \sum_{m=0}^M C_s^M(k, m, f') \cdot Y_{G,s}(k - m, f'), \quad (4)$$

where M denotes the order of the complex filter. This deep filtering stage is performed up to $f_{df} \approx 6$ kHz, where most of the background spectral content resides and where the blending of impulsive and stationary components occurs. Above f_{df} , only real-valued filtering is applied. This two-step approach reduces both memory and computational costs by minimizing the size of the complex filters required for source separation.

For further details on the model architecture we refer to the DeepFilterNet2 article [16], which is reproduced as is, with the only difference that the decoders' outputs are doubled for the prediction of each source as shown in Figure 1.

2.2. Loss functions

We adopt the same loss functions as described in [16] for reconstructing each source, i.e., the impulsive and stationary background

components, as well as for the mixture $\hat{Y}_i + \hat{Y}_s$. For each source Z to reconstruct, we compute the following spectrogram loss (\mathcal{L}_{SP}) between the predicted source \hat{Z} and the target Z ,

$$\mathcal{L}_{SP}(\hat{Z}, Z) = \|\hat{Z}^c - Z^c\|_2^2 + \|\hat{Z}^c e^{j\Phi_{\hat{Z}}} - Z^c e^{j\Phi_Z}\|_2^2, \quad (5)$$

where $\|\cdot\|_2$ denotes the l_2 -norm, $c = 0.6$ is a compression factor used to approximate perceived loudness [17] and Φ represents the phase. Additionally, a multi-resolution (MR) spectrogram loss is computed by converting Z back to the time domain using an inverse short-time fourier transform (ISTFT), followed by multiple STFTs with different window sizes of length $\{6, 12, 23\}$ ms:

$$\mathcal{L}_{MR}(\hat{Z}, Z) = \sum_m \|\hat{Z}_m^c - Z_m^c\|_2^2 + \sum_m \|\hat{Z}_m^c e^{j\Phi_{\hat{Z}}} - Z_m^c e^{j\Phi_Z}\|_2^2, \quad (6)$$

where m indexes the window sizes. The overall loss for the element Z is then given by

$$\mathcal{L}(\hat{Z}, Z) = \lambda_{SP} \mathcal{L}_{SP}(\hat{Z}, Z) + \lambda_{MR} \mathcal{L}_{MR}(\hat{Z}, Z), \quad (7)$$

with $\lambda_{SP} = 1000$ and $\lambda_{MR} = 500$. Finally, the complete training loss sums the contributions of each source to be reconstructed and the mixture:

$$\mathcal{L}_{total} = \lambda_i \mathcal{L}(\hat{Y}_i, Y_i) + \lambda_s \mathcal{L}(\hat{Y}_s, Y_s) + \lambda_m \mathcal{L}(\hat{Y}_i + \hat{Y}_s, X). \quad (8)$$

where $\lambda_i = \lambda_m = 1$ and $\lambda_s = 10$. Since most of the stationary signal is easy to reconstruct—impulse-free regions require little to no filtering—the loss $\mathcal{L}(\hat{Y}_s, Y_s)$ tends to be dominated by these areas. Early experiments showed that applying a weight $\lambda_s > 1$ improves the reconstruction of short segments around impulses by giving them more weight. As for the loss on the mixture, this is calculated to ensure a certain cohesion between the predictions of the background and impulsive sounds, but it remains globally dominated by the other two terms.

3. DATA GENERATION PIPELINE

As mentioned earlier, the key step in the proposed system is the data generation process. We generated training, validation, and test datasets by combining stationary background acoustic scenes with clean impulsive sounds to replicate realistic acoustic environments. The background datasets selected are Dcase2018 Task 1 [18], Cas2023 [19], CochIScene [20], LitisRouen [21], and ARTE [22], which provide a wide variety of acoustic scenes and are commonly used in acoustic scene classification tasks. Additionally, we generated synthetic background scenes by augmenting pink noise with random equalization, gain transitions, reverberation, and the addition of low-level Gaussian noise to simulate stationary noises, such as ventilation noise.

For impulsive sounds, we used datasets containing isolated sound events: ESC50 [23], Nonspeech7k [24], ReaLISED [25], and Vocal-Sound [26]. We also included two datasets of one-shot percussive instruments: FreesoundOneShotPercussive [27] and other drum samples. To further increase the variety of impulsive sounds, we generated synthetic events from chirps, harmonic summation, and AR filtering of white noise modulated by asymmetric Gaussian envelopes. All the code for generating the synthetic sounds, both backgrounds and impulses, will be made publicly available.

3.1. Impulsive sounds

In this work, we define an *impulsive acoustic event* as a brief and isolated sound that perceptually stands out from the ambient background. An important aspect of this definition is the fact that the superposition or repetition of impulsive sounds over time, which form a distinct sound layer (e.g., applause or rain), are not considered as impulsive acoustic events. For example, we differentiate between an isolated hammer blow (impulsive) and a continuous burst of jackhammer blows over several seconds (texture to remain in the background component of our model).

3.2. Dataset Pre-processing

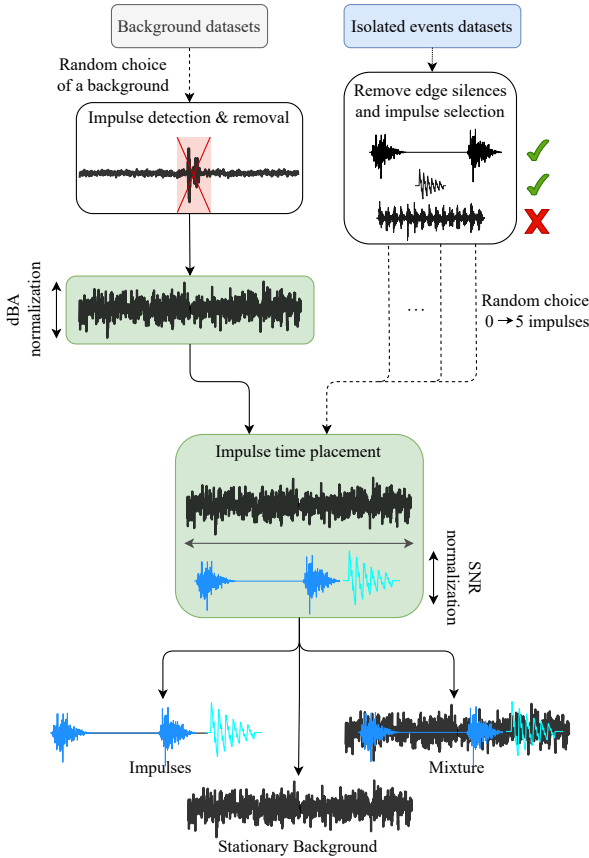


Fig. 2: Data generation pipeline.

To generate suitable training data, it is essential to have *clean* background datasets (i.e. free from discernible impulsive sounds) and clean sounds that are genuinely impulsive. Consequently, all datasets require pre-processing to remove unwanted elements.

For background datasets, we apply an impulse removal procedure at a reduced sampling rate of 16 kHz. First, we detect onsets using *librosa* with a hop length of 512 samples and a delta threshold of

20% of the signal’s maximum amplitude. Then, a Gabor decomposition is applied to a 5-second window around each detected onset using multi-Gabor dictionaries [28] with various temporal supports ($N_w = 32, 64, 128, 512$ ms). to obtain the atoms coefficients $c_w(n_w, f_w)$, n_w and f_w being the time and frequency bins. An impulsive event is characterized by a localized burst of energy across frequencies. To identify this pattern in the coefficients obtained, we sum the frequency contributions at each time step $\tilde{c}(n_w) = \sum_{f_w} c_w(n_w, f_w)$. For each window size, the coefficients are interpolated to the resolution of the smallest window, and the contributions from the 3 smallest windows are summed $\tilde{c}_{overall} = \tilde{c}_{w_1} + \tilde{c}_{w_2} + \tilde{c}_{w_3}$. A peak detection is performed on the obtained coefficients with a distance parameter of 100 ms, a prominence of 30% the maximum value of $\tilde{c}_{overall}$ and a height parameter equal to the maximum coefficient value c_{w_4} on the larger temporal support. If a detected peak is within 200 ms of the onset, the onset is validated as an impulsive sound and removed, with crossfading applied to reconnect the background segments.

For datasets containing isolated sound events, we filter out non-impulsive sounds by applying the following procedure: for each audio sample in the dataset, we first remove the silences at the edges; then, we retain only the sounds shorter than half a second. For longer sounds, we calculate the proportion of silence from the root mean square (RMS) envelope (using a 5% threshold of the 99th percentile of the envelope) and keep only those with a sufficiently high ratio of silence (50% for signals less than 1s, and 75% for longer ones).

3.3. Dataset generation

The generated dataset consists of 5-second acoustic scenes, sampled at $f_s = 44100$ Hz, each comprising a background selected randomly from the pre-processed background datasets and several impulsive events chosen from the pre-processed isolated sound datasets. The generation process is presented Figure 2.

To prevent producing a biased dataset and to avoid over- or under-representation of different acoustic scene categories and impulsive sound types, we organize and unify the various dataset labels, impulse events and backgrounds separately, into a taxonomy using the SALT framework [29]. When drawing background or impulsive sounds, they are selected from subsets of the datasets that contain the same number of samples for each class, with the exception of those with very few items. The background track, $y_b(t)$, is normalized to a dBA level sampled from a realistic distribution based on the scene label, while the impulsive sounds are placed randomly without overlapping along the time axis, normalized to reach randomly selected target signal-to-noise ration (SNR) levels, and grouped into a single track, $y_i(t)$. Augmentations are applied to the impulsive sounds (e.g., equalization, reverb, time stretching, and pitch shifting), and a final impulse response is applied to both the background and impulsive tracks. Each set of sources is then exported, as well as the mixture. In total, 50 hours of mixture data were generated for training, 20 hours for validation, and 10 hours for testing.

4. EVALUATION

We present here the evaluation framework used to assess the performance of our proposed method, detailing the training setup, baseline methods and the obtained results on the test set.

4.1. Training setup and baselines

IS³ was trained with the 50h dataset described in Section 3 for a maximum of 150 epochs with early stopping using a batch size of 32 and an Adam optimizer [30] with a learning rate of 10^{-3} . The audio parameters are the following: we use $N_{erb} = 24$ and $N_{feat} = 256$, just under 6 kHz, and an order $M = 8$ for the complex filter. The

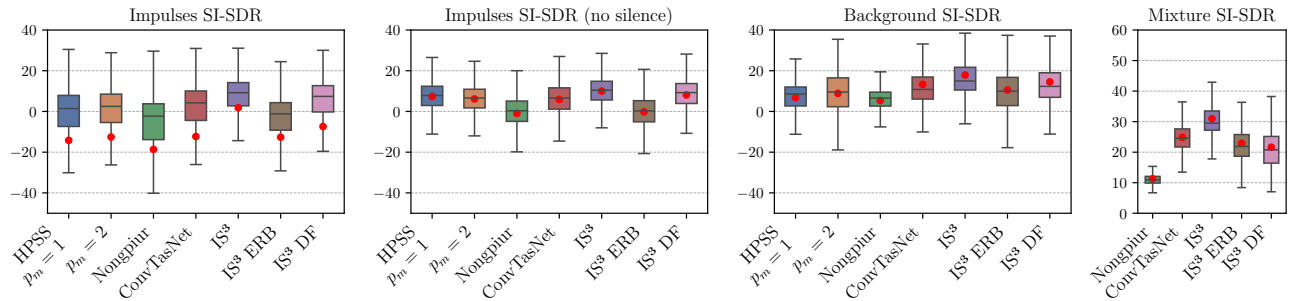


Fig. 3: Obtained SI-SDR for the impulses, the stationary backgrounds and the mixture, with the different baseline configurations and the proposed IS³ system. The red dots are the mean values.

result is a model with around 2.2 million parameters. Additionally we trained ERB-only (1.2 M parameters) and deep-filtering-only (1.4 M parameters with the same N_{feat}) variants of IS³ as ablation studies to demonstrate the benefits of the staged design.

We compare our approach with several baselines. First, we reimplement Nongpiur’s method [6] for impulsive noise removal in speech signals, which employs Daubechies wavelets to decompose the noisy signal. The method identifies impulses by applying median filtering to wavelet coefficients across scales and attenuates them to neighboring median levels. We adapt this method to our use case and sampling rate by modifying the wavelet order—choosing 13 instead of 6—and adjusting the factors that control the dynamic thresholds for detecting impulse coefficients: $k_s = 2$ for fine scales and $k_c = 1$ for coarser scales. Additionally, while the original approach only predicts a clean signal with attenuated impulses, we extend it to include an impulsive sound reconstruction stage by retaining only the wavelet coefficients identified as impulses. Secondly, we compare with the median-filtering HPSS masking method [13], specifically its `librosa` implementation [31], [32]. In this baseline, the percussive component is treated as the impulsive part, while the harmonic and residual components (if the margin parameter is greater than 1 are considered as the stationary background. We assess different HPSS configurations by varying the margin parameter, or separation factor, p_m . Finally, we re-trained a Conv-TasNet model [33], using `asteroid` [34], adapted to 44.1 kHz input [35] (6.3M parameters) to compare with a time-domain neural architecture. All the code for the data generation, the baselines as well as IS³ are available on [github](https://github.com/ClementineBerger/IS3)¹.

4.2. Results and discussion

The evaluation was conducted using the previously described test set. We use the scale-invariant signal-distortion-ratio (SI-SDR) metric [36] on the separated impulsive and stationary background components, as well as on the reconstructed mixtures for Nongpiur’s method and IS³. Since HPSS is a masking-based method, its mixture reconstruction is inherently perfect. Additionally, for impulsive sounds, the SI-SDR is computed both with and without silent segments to assess not only reconstruction quality but also the model’s ability to preserve silences and prevent background leakage. Statistical significance between IS³ and each baseline configuration is evaluated using the Wilcoxon test over 100 batches of 50 samples, with Bonferroni correction. Results are shown in Figure 3 and the p -values are presented in Table 1.

IS³ consistently achieves higher SI-SDR scores for both impulses and backgrounds. The small gap between SI-SDR scores with and without silence suggests that our method accurately reconstructs both impulses and interleaving silences. In contrast, HPSS and Nongpiur’s

Table 1: Statistical significance between the SI-SDR distributions obtained with the proposed IS³ model and the different baselines and ablations. p -values are evaluated using the Wilcoxon test over 100 batches of 50 samples, with Bonferroni correction.

	Impulse p -value		Background	Mixture
	Silence	No silence		
HPSS				
$p_m = 1$	$3.74 \cdot 10^{-2}$	0.17	$1.40 \cdot 10^{-5}$	—
$p_m = 2$	$5.27 \cdot 10^{-2}$	$6.02 \cdot 10^{-2}$	$2.66 \cdot 10^{-3}$	—
Nongpiur	$4.96 \cdot 10^{-3}$	$1.47 \cdot 10^{-5}$	$1.19 \cdot 10^{-7}$	$5.18 \cdot 10^{-10}$
Conv-TasNet	$1.75 \cdot 10^{-8}$	$6.72 \cdot 10^{-8}$	$8.62 \cdot 10^{-7}$	$1.60 \cdot 10^{-7}$
IS ³ ERB	$2.08 \cdot 10^{-8}$	$2.26 \cdot 10^{-9}$	$5.44 \cdot 10^{-9}$	$5.93 \cdot 10^{-8}$
IS ³ DF	$3.96 \cdot 10^{-7}$	$8.08 \cdot 10^{-6}$	$2.06 \cdot 10^{-5}$	$6.62 \cdot 10^{-9}$

method suffer significant SI-SDR degradation when silences are considered, indicating background leakage into the impulsive sound track. This leakage also lowers HPSS’s background reconstruction performance compared to IS³. Finally, while our approach does not strictly guarantee perfect mixture reconstruction like masking methods, it achieves a remarkably high SI-SDR on the mixture. Conv-TasNet and ablations achieve intermediate results demonstrating the value of the architecture chosen for IS³.

Finally, it is important to note that both signal processing baseline methods suffer from a reliance on a challenging and highly impulsive noise type dependant parameter selection. This dependency reduces their performance in our experiments, which involve a wide variety of impulsive sound types. In contrast, our approach offers superior generalization and eliminates the need for noise-specific parameter tuning. For a qualitative comparison, audio examples are provided in the supplementary materials² for both synthetic and real-world data.

5. CONCLUSIONS

In this paper, we have introduced IS³, a solution for Impulsive–Stationary Sound Separation, designed to isolate generic impulsive acoustic events from stationary backgrounds within an acoustic scene. Our approach leverages and adapts the DeepFilterNet2 two-stage deep filtering process for this task and is trained using a dedicated dataset generated through a sophisticated data generation pipeline to ensure diversity and balance across sound categories. Evaluation results demonstrate that IS³ is successful at separating both impulsive and stationary components while minimizing background leakage. These results demonstrate that a learning-based approach trained on well-designed data is well-suited for the task and can achieve strong performance even with a relatively lightweight model. In particular, the proposed approach outperforms the classic HPSS masking method and wavelet filtering by a large margin in terms of SI-SDR.

¹<https://github.com/ClementineBerger/IS3>

²<https://clementineberger.github.io/IS3/>

REFERENCES

- [1] Z. Brajević, “Elimination of unwanted signals in audio materials using wavelet transform,” in *Proceedings ELMAR-2011*, 2011, pp. 229–233.
- [2] M. Ruhlmann, J. Bitzer, M. Brandt, and S. Goetze, “Reduction of Gaussian, supergaussian, and impulsive noise by interpolation of the binary mask residual,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1680–1691, 2015.
- [3] S. J. Godsill and P. J. Rayner, “Robust treatment of impulsive noise in speech and audio signals,” *Lecture Notes-Monograph Series*, pp. 331–342, 1996.
- [4] A. Subramanya, M. L. Seltzer, and A. Acero, “Automatic removal of typed keystrokes from speech signals,” *IEEE signal processing letters*, vol. 14, no. 5, pp. 363–366, 2007.
- [5] A. Sugiyama, “Single-channel impact-noise suppression with no auxiliary information for its detection,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 127–130.
- [6] R. C. Nongpiur, “Impulse noise removal in speech using wavelets,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 1593–1596.
- [7] A. Sugiyama, R. Miyahara, and K. Park, “Impact-noise suppression with phase-based detection,” in *21st European Signal Processing Conference (EUSIPCO 2013)*, 2013, pp. 1–5.
- [8] A. Medda and D. A. Alvord, “Separation of impulsive transients from broadband noise by wavelet filtering,” in *23rd AIAA/CEAS Aeroacoustics Conference*, 2017, p. 3192.
- [9] J. Wodecki, A. Michalak, R. Zimroz, T. Barszcz, and A. Wyłomańska, “Impulsive source separation using combination of Nonnegative Matrix Factorization of bi-frequency map, spatial denoising and Monte Carlo simulation,” *Mechanical Systems and Signal Processing*, vol. 127, pp. 89–101, 2019.
- [10] J. Young, A. Høst-Madsen, and E.-M. Nosal, “Impulsive source separation with application to sperm whale clicks,” in *2013 IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, 2013, pp. 147–152.
- [11] X. Wen, Y.-Y. Shi, and B. She, “Separation of impulsive acoustical events,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2004, pp. ii–733.
- [12] M. Sahmoudi, K. Abed-Meraim, and M. Benidir, “Blind separation of impulsive alpha-stable sources using minimum dispersion criterion,” *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 281–284, 2005.
- [13] D. Fitzgerald, “Harmonic/percussive separation using median filtering,” *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, 2010.
- [14] W. Mack and E. A. Habets, “Deep filtering: Signal extraction and reconstruction using complex time-frequency filters,” *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.
- [15] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, “Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7407–7411.
- [16] H. Schröter, A. Maier, A. N. Escalante-B, and T. Rosenkranz, “Deepfilternet2: Towards real-time speech enhancement on embedded devices for full-band audio,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.
- [17] J.-M. Valin, S. Tenneti, K. Helwani, U. Isik, and A. Krishnaswamy, “Low-complexity, real-time joint neural echo control and speech enhancement based on perceptnet,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7133–7137.
- [18] A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13.
- [19] J. Bai, M. Wang, H. Liu, H. Yin, Y. Jia, S. Huang, Y. Du, D. Zhang, D. Shi, W.-S. Gan, M. D. Plumley, S. Rahardja, B. Xiang, and J. Chen, “Description on IEEE ICME 2024 Grand Challenge: Semi-supervised Acoustic Scene Classification under Domain Shift,” 2024.
- [20] I.-Y. Jeong and J. Park, “Cochlscene: Acquisition of acoustic scene data using crowdsourcing,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 17–21.
- [21] A. Rakotomamonjy and G. Gasso, “Histogram of gradients of time-frequency representations for audio scene classification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2014.
- [22] A. Weisser, J. M. Buchholz, C. Oreinos, J. Badajoz-Davila, J. Galloway, T. Beechey, and G. Keidser, “The Ambisonic Recordings of Typical Environments (ARTE) Database,” *Acta Acustica United With Acustica*, vol. 105, no. 4, pp. 695–713, 2019.
- [23] K. J. Piczak, “Esc: Dataset for environmental sound classification,” in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.
- [24] M. M. Rashid, G. Li, and C. Du, “Nonspeech7k dataset: Classification and analysis of human non-speech sound,” *IET Signal Processing*, vol. 17, no. 6, p. e12233, 2023.
- [25] I. Mohino-Herranz, J. García-Gómez, M. Aguilar-Ortega, M. Utrilla-Manso, R. Gil-Pita, and M. Rosa-Zurera, “Introducing the realised dataset for sound event classification,” *Electronics*, vol. 11, no. 12, p. 1811, 2022.
- [26] Y. Gong, J. Yu, and J. Glass, “Vocalsound: A dataset for improving human vocal sounds recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 151–155.
- [27] A. Ramires, P. Chandna, X. Favory, E. Gómez, and X. Serra, “Neural percussive synthesis parameterised by high-level timbral features,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 786–790.
- [28] Z. Prusa, N. Holighaus, and P. Balazs, “Fast matching pursuit with multi-Gabor dictionaries,” *ACM Transactions on Mathematical Software (TOMS)*, vol. 47, no. 3, pp. 1–20, 2021.
- [29] P. Stamatiadis, M. Olvera, and S. Essid, “Salt: Standardized audio event label taxonomy,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)*, Tokyo, Japan, October 2024, pp. 161–165.
- [30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.
- [31] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, vol. 8, 2015.
- [32] J. Driedger, M. Müller, and S. Disch, “Extending harmonic-percussive separation of audio signals,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 611–616.
- [33] 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*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [34] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, “Asteroid: the PyTorch-based audio source separation toolkit for researchers,” in *Proc. Interspeech*, 2020.
- [35] A. Défossez, N. Usunier, L. Bottou, and F. Bach, “Music source separation in the waveform domain,” *arXiv preprint arXiv:1911.13254*, 2019.
- [36] J. Le 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.