SRPOL SUBMISSION TO DCASE 2024 CHALLENGE TASK 9: MODELING REAL AND IMAGINARY COMPONENTS, MIXIT AND SDR BASED LOSS

Technical Report

Michal Romaniuk[*]*, Justyna Krzywdziak*[*]

Samsung R&D Institute Poland

ABSTRACT

We present our solution to the DCASE 2024 challenge task 9 (Language-Queried Audio Source Separation). Our solution is based on the official baseline, with training dataset including FSD50k, Clotho and additionally extended with AudioCaps. We show that the additional data improve results throughout the training process. We explore changing the ratio masking method from spectrogram amplitude and phase to individual masks for real and imaginary components. We also investigate how different losses, such as Mixit loss and SDR based loss, affect the training process.

***Index Terms*—** DCASE, audio source separation, language-queried audio source separation

1. INTRODUCTION

Recent developments in image-text embedding architectures such as CLIP [1] and also audio-text embedding architectures [2] have enabled novel approaches to the problem of audio source separation [3, 4, 5]. Among previous methods, some have limited the problem to a set of predefined sound classes [6]. Some have focused particularly on speech [7, 8, 9] or musical instruments [10, 11]. Others proposed to simply separate a recording into a fixed number of sources [12], which leaves their classification to downstream processing. Still others relied on audio-visual supervision [13], or providing example instances of the audio class to extract from mixture [14].

In contrast, language-queried methods employ audio-text alignment models to produce embeddings from text prompts that are then fed to the audio separation model [3, 4, 5]. This approach is more practical for several reasons. First, it is more flexible than predefined labels and second, text prompts are easier to generate than visual prompts (for audio-visual models) or audio prompts (audio prompted models). Text queries can also be combined with video [15] or other features [16].

1.1. Evaluation submissions

We submitted outputs from 4 systems to the DCASE evaluation server:

SRPOL system 1 This model takes real and imaginary spectrogram components as input and produces individual masks for these components.

SRPOL system 2 This is the baseline, except trained for 1.5 million steps with the original 10^{-3} learning rate, followed by 100 thousand steps with 10^{-4} learning rate.

SRPOL system 3 This is the baseline, except trained for 1.084 million steps with the original 10^{-3} learning rate.

SRPOL system 4 This is the baseline, except trained for 1.3 million steps with the original 10^{-3} learning rate followed by 40 thousand steps with 10^{-5} learning rate.

1.2. Contributions

Our contributions include:

1. A comparison of training on different datasets: FSD50k with auto-generated captions, Clotho and AudioCaps
2. A comparison of spectrogram magnitude masking and complex ratio masking
3. A comparison of losses used in training - baseline l1 loss, MixIT loss and SDR based loss

2. DATASETS

2.1. FSD50k

The FSD50k dataset [17] consists of 51197 audio clips downloaded from Freesound, which were manually labeled with 200 classes from the AudioSet [18] ontology. The total audio duration in FSD50k is over 108 hours.

Challenge organizers provide text captions for FSD50k that were auto-generated with ChatGPT from the original FSD50k labels. For training we used the captions from the challenge dev subset (fsd50k_dev_auto_caption.json), mapping to 40966 files with over 80 hours of audio. For validation we use the method from the challenge evaluation script with the audio mixing recipes provided in the lass_synthetic_validation.csv file.

2.2. Clotho

The original Clotho dataset [19] comprises 4981 audio clips, each lasting between 15 to 30 seconds. The audio samples are sourced from the Freesound platform, consisting of acoustic environments and sound events. These clips encompass a spectrum of everyday sounds, including natural environments, human activities, musical instruments, mechanical noises, and more. Each audio clip is annotated with five different captions, ranging from 8 to 20 words in length.

The captions were generated using a crowdsourcing approach, involving annotators who were native English speakers and had undergone specific training to ensure high-quality descriptions. Each

[*]Equal contribution

annotator listened to the audio clips and provided textual descriptions that accurately captured the auditory content. The annotations were subsequently reviewed and refined to maintain consistency and precision across the dataset.

Since the original Clotho v1.0, the authors released versions 2.0 and 2.1. In our experiments we used version 2.1 which consists of 5929 audio clips, with a total duration of over 37 hours. For training we used the "dev" subset, which consists of 3839 files with total audio duration of nearly 24 hours.

2.3. AudioCaps

The AudioCaps dataset [20] contains 46,000 audio clips sourced from the AudioSet dataset [18], each ranging from 5 to 10 seconds in duration. These clips cover a wide variety of sound events and environments, including natural sounds, urban noises, human activities, music, and more, ensuring comprehensive coverage of everyday audio phenomena. Each audio clip is annotated with a detailed caption describing the sound events present in the clip that range from a few words to complete sentences.

The annotations were generated through a combination of human and automated processes. Initially, a subset of audio clips was manually annotated by trained workers who listened to the clips and provided descriptive captions. These human-generated captions served as a basis for training a machine learning model, which was subsequently used to generate captions for the remaining clips. This hybrid approach leverages the accuracy of human annotation and the scalability of automated processes.

Since AudioCaps is based on AudioSet, there is the problem of clips that become unavailable over time. Overall, we managed to build a dataset of 50166 files with total audio length of over 137 hours. For training we used 43703 audio files with total duration of nearly 120 hours.

3. SYSTEM DESCRIPTION

3.1. Baseline system

The baseline system, based on AudioSep [4], was provided by the challenge organizers. It combines a ResUNet audio separation model with a CLAP audio-text embedding model, which is used to provide conditioning to ResUNet. There is also a checkpoint provided, trained on the challenge development data set.

3.2. Real and imaginary ratio masking

We test replacing the original method of modeling amplitude and phase as separate variables with modeling the real and imaginary components of the STFT representation. More precisely, while the baseline solution is predicting a spectrogram amplitude ratio mask and a phase correction mask, our model directly estimates ratio masks for the real and imaginary components:

$$\widehat{X} = M_{Re} \odot Re\,(Y) + iM_{Im} \odot Im\,(Y) \tag{1}$$

where Y is the mixture STFT, $\widehat{X}$ is the estimated separated audio STFT, M_{Re} and M_{Im} are the real and imaginary ratio masks.

This approach is fairly common in speech enhancement, although instead of ratio masks for real and imaginary components, the model outputs are typically interpreted as real and imaginary components of a complex-valued mask, which is then complex-multiplied with the mixture STFT [21]:

$$\begin{aligned}\widehat{X} = {} & Re\,(M) \odot Re\,(Y) - Im\,(M) \odot Im\,(Y) \\ & + i\left(Re\,(M) \odot Im\,(Y) + Im\,(M) \odot Re\,(Y)\right)\end{aligned} \tag{2}$$

This is often referred to as complex ratio masking (CRM). However, our model uses the masking definition in (1) for its simplicity.

Additionally, our model takes the real and imaginary components of the mixture STFT as input features, rather than the complex magnitude that was used in the baseline.

3.3. MixIT Loss

MixIT [22] Loss is a training objective designed to address the challenges of audio source separation in unsupervised learning scenarios. Unlike traditional methods that require clean, isolated sources for training, MixIT Loss enables models to learn directly from mixed audio inputs, making it suitable for real-world applications where obtaining clean source data is impractical.

The MixIT framework operates on the principle of mixing multiple audio sources to create training mixtures. The key steps in the methodology include:

1. Input Mixtures: Two mixtures of audio sources, x_1 and x_2 are combined to form a new mixture $x_{mix} = x_1 + x_2$
2. Separation: A neural network is trained to separate x_{mix} into a set of components, aiming to approximate the original sources from x_1 and x_2
3. Permutation Invariant Training: The separated components are matched to the original sources in a permutation-invariant manner, which means the order of the separated sources does not need to match the order of the input sources.

The MixIT Loss is calculated by evaluating how well the separated components can reconstruct the original input mixtures. For each separated component, a reconstruction of the original mixtures x_1 and x_2 is created. The loss is computed as the sum of reconstruction errors between the original mixtures and their reconstructions from the separated components. Equation 3 represents the mathematical expression of the loss where $\hat{s}_j$ are the separated components, N is the number of separated components, and π_{ij} represents the permutation matrix that aligns separated components to the original mixtures. This ensures that the network learns to produce components that can be recombined to approximate the original mixtures accurately.

$$MixITLoss = min_{\pi}(\sum_{i=1}^{2} \|x_i - \sum_{j=1}^{N} \pi_{ij}\hat{s}_j\|^2) \tag{3}$$

3.4. SDR based loss

Inspired by [23] that presents a modified version of the CLAP network, we utilized the loss used in this paper for our system. The training loss is defined as a combination of negative signalto-distortion ratio (SDR) and negative scale-invariant signal-to-distortion ratio (SISDR) as shown in equation 4.

$$SDRLoss = -\lambda SDR(\hat{x}, x) - (1-\lambda)SISDR(\hat{x}, x), \tag{4}$$

where

$$SDR(\hat{x}, x) = 10log_{10}(\frac{\|x\|^2}{\|x - \hat{x}\|^2}), \quad (5)$$

$$SISDR(\hat{x}, x) = 10log_{10}(\frac{\|\frac{\|\hat{x}^T x\|^2}{\|x\|^2} x\|^2}{\|\frac{\|\hat{x}^T x\|^2}{\|x\|^2} x - \hat{x}\|^2}), \quad (6)$$

where $\hat{x}$ and x denote the estimated waveform and the ground truth waveform. The λ parameter was set to 0.9 in our experiments.

4. EXPERIMENTS

We use the official baseline code for training all models. The optimization algorithm is Adam (AMSgrad version) [24, 25], with learning rate set to 10^{-3} and with 10000 warm-up steps (except for fine-tuning where we reduced the learning rate and switched off warm-up). Weight decay is set to zero.

4.1. Combining datasets

The goal of this set of experiments was to determine if we could improve the results by adding more data. First we trained the baseline system on FSD50k data, using the captions provided by the challenge organizers. Next, we extended the training dataset with Clotho, and then we also added AudioCaps. The model trained on all three datasets combined was later additionally fine-tuned on the same data with learning rate reduced to 10^{-4}.

4.2. Real and imaginary ratio masks

We compare the proposed method of modeling real and imaginary components to the original baseline method of modeling amplitude and phase.

4.3. MixIT and SDR based loss

The baseline L1 Loss, also known as Mean Absolute Error (MAE), measures the average absolute difference between predicted and actual signal values. It is less sensitive to extreme values than L2 Loss (Mean Squared Error), making it more robust to noise in the data but it necessitates clean, isolated signals as training data, limiting its applicability in scenarios with unlabelled or mixed data. Our goal was to find better-suited loss for this task. We chose MixIT Loss because it allows training models based on mixtures of sounds without needing access to clean source signals and SDR-Based Loss which measures the quality of source separation based on the signal-to-distortion ratio, considering both amplitude and phase of the signal.

5. RESULTS

5.1. Combining datasets

We present the evolution of validation SDR and SI-SDR throughout the training process in Fig 1 and Fig 2.

The model based only on FSD50k has little improvement in SDR past 500k training steps while also declining in SI-SDR past this point.

With the addition of Clotho, SDR and SI-SDR grow up to 1.4 million steps of training. Both metrics are higher throughout the training process than with FSD50k alone and the trained model is also substantially better.

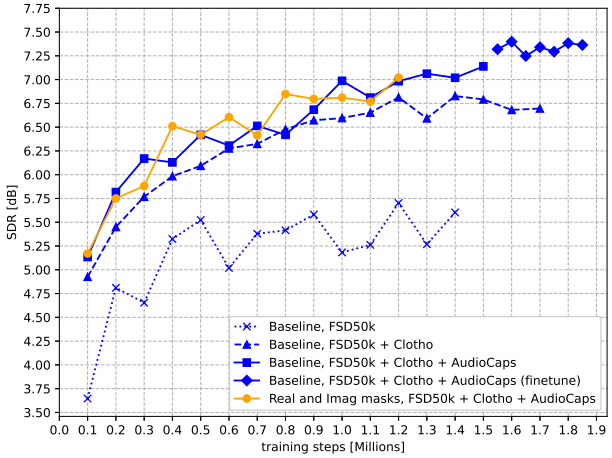

Figure 1: Challenge FSD50k validation set SDR

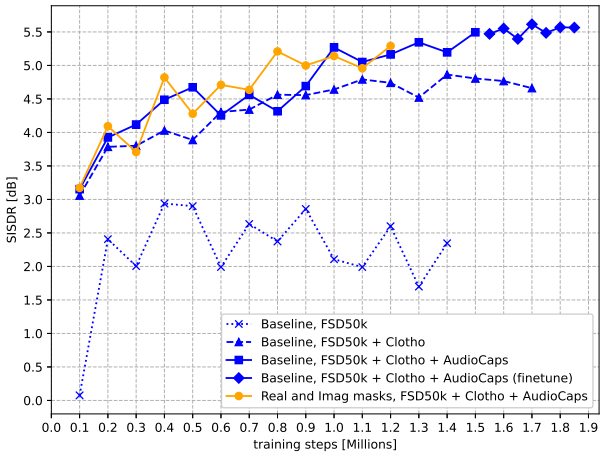

Figure 2: Challenge FSD50k validation set SI-SDR

After expanding the training dataset further with AudioCaps, both SDR and SI-SDR were still increasing at 1.5 million steps, which is the farthest point that we reached while training this model with the default learning rate. This model was further fine-tuned with reduced learning rate (10^{-4}), which further boosted SDR and SI-SDR, although it is possible that it could have been improved with the original learning rate.

The fine-tuned checkpoint at 1.6 million steps (which had the best validation SDR in this run) was used to generate the submission referred to as SRPOL system 2.

5.2. Real and imaginary ratio masks

The results of this experiment are also presented in Fig 1 and Fig 2. We only managed to train this model up to 1.2 million steps and up to this point the results of this model were close to the ones attained by the baseline trained on the same data.

The checkpoint at 1.2 million steps (which had the best valida-

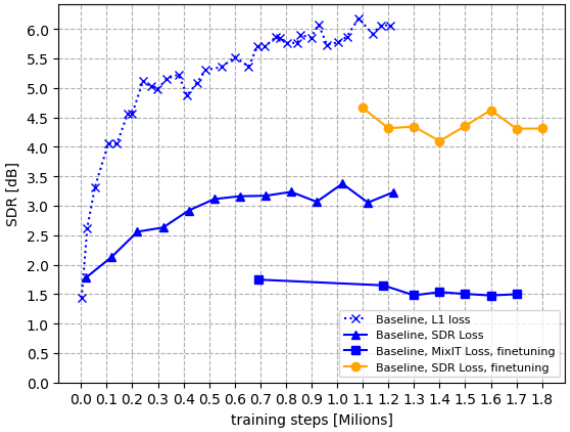

Figure 3: Challenge FSD50k validation set SDR

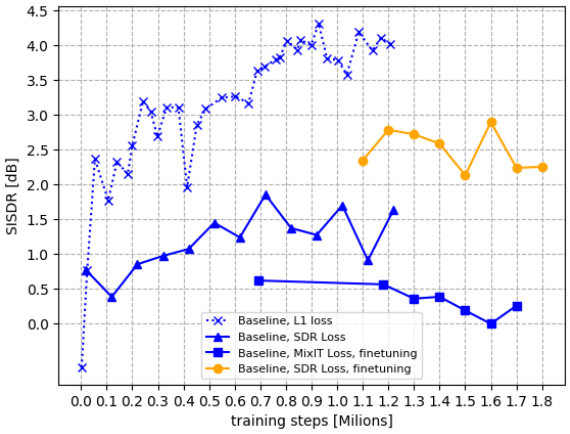

Figure 4: Challenge FSD50k validation set SISDR

tion SDR and SI-SDR in this run) was used to generate the submission referred to as SRPOL system 1.

5.3. MixIT and SDR based loss

Results of the loss experiments are presented in Fig 3 and Fig 4. Each model is trained on the extended dataset (FSD50K, Clotho, and AudioCaps). The best model is the baseline trained from scratch on an extended dataset with L1 loss and this is the SRPOL system 3. SRPOL system 4 is the same model but finetuned from step 1300000 with L1 loss and learning rate 10^{-5}. MixIT loss achieves the worst results. SDR-based loss is better but still, it does not achieve baseline results even after more than a million steps. When fine-tuning the baseline model with SDR-based loss the results are also not satisfactory.

5.4. Submitted models

The validation results for the models whose outputs were submitted for evaluation are shown in Table 1.

Table 1: Validation results for the models submitted for evaluation

Model	SDR	SDRi	SI-SDR
Real and Imag masks, FSD50k + Clotho + AudioCaps (*SRPOL system 1*)	7.021	6.986	5.291
Baseline, FSD50k + Clotho + AudioCaps (finetune) (*SRPOL system 2*)	7.398	7.363	5.551
Baseline, FSD50k + Clotho + AudioCaps (*SRPOL system 3*)	6.181	6.146	4.188
Baseline, FSD50k + Clotho + AudioCaps (finetune) (*SRPOL system 4*)	6.282	6.247	4.620

6. CONCLUSIONS

Our main conclusion is that the training of language-queried audio source separation models benefits from extending the training data set. Future work may explore expanding it further still. Another possible direction is to explore the effect of implementing complex ratio masking exactly (as opposed to separate masks for real and imaginary components, as implemented in our experiments). Furthermore, future experiments can be extended to include additional types of loss functions or combinations thereof. Each loss function, whether L1 Loss, MixIT Loss, or SDR-Based Loss, has its strengths and limitations. Exploring new or hybrid loss functions can leverage these strengths and mitigate the limitations, potentially leading to more robust and efficient training methods. Combining different loss functions could enhance model performance by providing a more comprehensive optimization and improving generalization to real-world scenarios. Such explorations could lead to the development of more sophisticated models that perform better in diverse and challenging acoustic environments.

7. ACKNOWLEDGMENT

The preparation of this report was partially aided by ChatGPT (`https://openai.com/chatgpt/`).

8. REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate what you describe: Language-queried audio source separation," in *Proc. INTERSPEECH 2022*, 2022.

[4] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *arXiv preprint arXiv:2308.05037*, 2023.

[5] K. Kilgour, B. Gfeller, Q. Huang, A. Jansen, S. Wisdom, and M. Tagliasacchi, "Text-driven separation of arbitrary sounds," in *"INTERSPEECH 2022"*, 2022, pp. 5403–5407.

[6] Q. Kong, K. Chen, H. Liu, X. Du, T. Berg-Kirkpatrick, S. Dubnov, and M. D. Plumbley, "Universal source separation with weakly labelled data," *arXiv preprint arXiv:2305.07447*, 2023.

[7] L. Liu, H. Guan, J. Ma, W. Dai, G. Wang, and S. Ding, "A mask free neural network for monaural speech enhancement," in *"INTERSPEECH 2023"*, 2023, pp. 2468–2472.

[8] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6648–6652.

[9] 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.

[10] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

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

[13] E. Tzinis, S. Wisdom, T. Remez, and J. R. Hershey, "Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation," in *European Conference on Computer Vision*. Springer, 2022, pp. 368–385.

[14] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Zero-shot audio source separation through query-based learning from weakly-labeled data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, 2022, pp. 4441–4449.

[15] R. Tan, A. Ray, A. Burns, B. A. Plummer, J. Salamon, O. Nieto, B. Russell, and K. Saenko, "Language-guided audio-visual source separation via trimodal consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 575–10 584.

[16] E. Tzinis, G. Wichern, P. Smaragdis, and J. Le Roux, "Optimal condition training for target source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[18] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[19] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[20] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[21] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[22] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.

[23] H. Ma, Z. Peng, M. Shao, J. Liu, X. Li, and X. Wu, "Clapsep: Leveraging contrastive pre-trained models for multi-modal query-conditioned target sound extraction," *arXiv preprint arXiv:2402.17455*, 2024.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," *arXiv preprint arXiv:1904.09237*, 2019.