Leveraging Audio-Only Data for Text-Queried Target Sound Extraction

Saijo, Kohei; Ebbers, Janek; Germain, François G; Khurana, Sameer; Wichern, Gordon; Le Roux, Jonathan

TR2025-033 March 08, 2025

Abstract

The goal of text-queried target sound extraction (TSE) is to extract from a mixture a sound source specified with a natural- language caption. While it is preferable to have access to large-scale text-audio pairs to address a variety of text queries, the limited number of available high-quality text-audio pairs hinders the data scaling. To this end, this work explores how to leverage audio-only data without any captions for the text-queried TSE task to potentially scale up the data amount. A straightforward way to do so is to use a joint audio-text embedding model, such as the contrastive language-audio pre-training (CLAP) model, as a query encoder and train a TSE model using audio embeddings obtained from the ground-truth audio. The TSE model can then accept text queries at inference time by switching to the text encoder. While this approach should work if the audio and text embedding spaces in CLAP were well aligned, in practice, the embeddings have domain-specific information that causes the TSE model to overfit to audio queries. We investigate several methods to avoid overfitting and show that simple embedding-manipulation methods such as dropout can effectively alleviate this issue. Extensive experiments demonstrate that using audio-only data with embedding dropout is as effective as using text captions during training, and audio-only data can be effectively leveraged to improve text-queried TSE models.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Leveraging Audio-Only Data for Text-Queried Target Sound Extraction

Kohei Saijo[1,2], *Janek Ebbers*[1], *François G. Germain*[1], *Sameer Khurana*[1], *Gordon Wichern*[1], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA [2]Waseda University, Tokyo, Japan

***Abstract*—The goal of text-queried target sound extraction (TSE) is to extract from a mixture a sound source specified with a natural-language caption. While it is preferable to have access to large-scale text-audio pairs to address a variety of text queries, the limited number of available high-quality text-audio pairs hinders the data scaling. To this end, this work explores how to leverage audio-only data without any captions for the text-queried TSE task to potentially scale up the data amount. A straightforward way to do so is to use a joint audio-text embedding model, such as the contrastive language-audio pre-training (CLAP) model, as a query encoder and train a TSE model using audio embeddings obtained from the ground-truth audio. The TSE model can then accept text queries at inference time by switching to the text encoder. While this approach should work if the audio and text embedding spaces in CLAP were well aligned, in practice, the embeddings have domain-specific information that causes the TSE model to overfit to audio queries. We investigate several methods to avoid overfitting and show that simple embedding-manipulation methods such as dropout can effectively alleviate this issue. Extensive experiments demonstrate that using audio-only data with embedding dropout is as effective as using text captions during training, and audio-only data can be effectively leveraged to improve text-queried TSE models.**

***Index Terms*—Text-queried target sound extraction, CLAP, embedding dropout**

I. Introduction

Audio source separation is a fundamental technique for computational scene analysis. High-fidelity separation has been achieved using neural network (NN)-based approaches [1]–[4], pioneered by deep clustering [5] and permutation invariant training [5], [6]. While source separation separates all the sources in a mixture, another line of approaches, target sound extraction (TSE), aims to extract only a target source specified by a query, such as a speaker ID [7], [8] or class labels [9]. More recently, text-queried TSE [10], [11] has been attracting more attention due to its high versatility.

While text-queried TSE has the potential to work on any classes of audios as it accepts natural language as a query, realizing this in practice remains a challenge because it would require large-scale high-quality paired text-audio data from many domains. As demonstrated in [12], a text-queried TSE model trained on a small-scale dataset [10] does not generalize to out-of-domain data. In contrast, AudioSep [12] achieved better generalization than [10] by using larger-scale datasets. Still, results in [13] imply that there is room for improvement by leveraging more data.

Several works have attempted to increase the amount of available data for text-queried TSE. In [13], increasing caption variety using a large language model (LLM)-based caption augmentation has been shown to be effective. CLIPSep [14] has been proposed to train a text-queried TSE model using image-audio pairs extracted from videos instead of text-audio pairs, by utilizing a contrastive language-image pre-training (CLIP) model [15] as the query encoder. The model is trained using images as queries while the text caption is used during inference, which enables the model to be trained without any text captions. However, using image queries to scale up the data amount may be ultimately ill-suited to text-queried TSE due to their intrinsic limitations in capturing out-of-screen and/or background sounds. Decent performance is achieved at the cost of a complicated training pipeline to estimate such out-of-screen and/or background sounds.

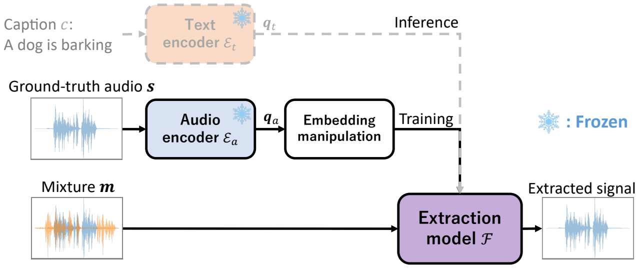
Caption c:
A dog is barking
Text encoder $\mathcal{E}_t$
q_t
Inference
Ground-truth audio $\boldsymbol{s}$
Audio encoder $\mathcal{E}_a$
$\boldsymbol{q}_a$
Embedding manipulation
Training
❄ : Frozen
Mixture $\boldsymbol{m}$
Extraction model $\mathcal{F}$
Extracted signal $\hat{\boldsymbol{s}}$

Fig. 1. Illustration of text-queried-TSE training with audio-only data. Audio embeddings extracted from ground-truth audio by a CLAP audio encoder are used to condition the extraction model during training, while text embeddings are used during inference. To prevent the extraction model from overfitting to audio-embedding-specific features, we apply a simple manipulation (e.g., dimension dropout, PCA, etc.) to the embeddings.

To improve the generalizability of text-queried TSE models without being limited by the scarcity and quality of text-audio pairs, a natural approach would be to rely on audio-only data without captions. Inspired by CLIPSep, one can think of using a contrastive language-audio pre-training (CLAP) [16] model as the query encoder so that audio embeddings aligned with corresponding text embeddings can be obtained during training. Although CLAP itself needs a large-scale text-audio pairs, training data can be further scaled up for text-queried TSE if we can leverage audio-only data. However, including audio-only data in such a way on top of paired text-audio data during training has not been explored so far. In [12], an extreme case, where only audio data is used for training, has been investigated but the training failed. This is likely because of the so-called *modality gap* phenomenon: the two embedding spaces in the CLAP models may not be very well aligned [17].

Based on the above observation, this work aims to develop ways to effectively leverage audio-only data for text-queried TSE without overfitting to CLAP audio embeddings. We investigate several methods for manipulating the CLAP audio embeddings, for example, by randomly dropping out dimensions, and show that this can significantly mitigate overfitting. The experiments demonstrate that audio-queried training with embedding dropout is as effective as normal text-queried training, even for the data that is not used for the CLAP training, which implies that we can potentially scale up the data amount for text-queried TSE. In addition, we show that such embedding manipulations are effective even in text-queried training and make the TSE model more robust to out-of-domain data.

This work was performed while K. Saijo was an intern at MERL.

II. Audio-queried training for text-queried TSE

A. Problem setup

Let us denote the mixture, target source, and non-target source as $\boldsymbol{m}$, $\boldsymbol{s}$, and $\boldsymbol{n} \in \mathbb{R}^L$, respectively, where $\boldsymbol{m} = \boldsymbol{s} + \boldsymbol{n}$ and L is the number of samples in the time domain. The goal of text-queried TSE is to extract $\boldsymbol{s}$ using a text caption c that represents the content of $\boldsymbol{s}$. Typically, the caption c is encoded into a D-dimensional text embedding $\boldsymbol{q}_t \in \mathbb{R}^D$ with a pre-trained text encoder $\mathcal{E}_t$, and $\boldsymbol{q}_t$ is given to a sound extraction model $\mathcal{F}$ as a conditioning vector:

$$\hat{\boldsymbol{s}} = \mathcal{F}(\boldsymbol{m}, \boldsymbol{q}_t), \quad \boldsymbol{q}_t = \mathcal{E}_t(c). \tag{1}$$

To scale up the training data for the text-queried TSE task, we are interested in methods that allow us to effectively incorporate audio-only data during training. To achieve that goal, we use CLAP [16] to get paired text and audio query encoders $\mathcal{E}_t$ and $\mathcal{E}_a$ (respectively), trained to learn a joint embedding space between text and audio. As illustrated in Fig. 1, during training, we use audio embeddings $\boldsymbol{q}_a \in \mathbb{R}^D$ obtained by inputting the *ground-truth target source* $\boldsymbol{s}$ to the CLAP audio encoder $\mathcal{E}_a$, instead of $\boldsymbol{q}_t$:

$$\hat{\boldsymbol{s}} = \mathcal{F}(\boldsymbol{m}, \boldsymbol{q}_a), \quad \boldsymbol{q}_a = \mathcal{E}_a(\boldsymbol{s}). \tag{2}$$

Such an audio-queried training could in principle work well if the text and audio embedding spaces in the CLAP model were well aligned (i.e., $\boldsymbol{q}_t \approx \boldsymbol{q}_a$ for pairs of a source $\boldsymbol{s}$ and associated caption c). However, in practice, the TSE model overfits to the CLAP audio embedding $\boldsymbol{q}_a$ because of the modality gap in audio and text subspaces, as detailed below.

B. Modality gap in CLAP

Although CLAP aims to train text and audio encoders to project their inputs to a shared embedding space, it has been shown that the training objective, particularly a low temperature in the contrastive loss, can lead to a *modality gap* [17] between the two embeddings subspaces. Prior work on CLAP [16], [18], [19] does use a low temperature value (e.g., initialized to 0.007), which suggests that existing CLAP models are likely to suffer from the modality gap. In addition, intuitively, audio embeddings may have richer information than text ones as text captions do not describe all the content of audios in most cases, which may also cause the gap. Investigating the average cosine similarity between text and audio embeddings from the LAION-CLAP model [18] and the Microsoft-CLAP model (MS-CLAP) [19] on the AudioCaps dataset [20], we found that they were indeed only around 0.4, even though AudioCaps is used for training.

C. Methods for alleviating modality gap

Based on the above discussion, it may be possible to train a CLAP model with a lesser modality gap by using a high temperature. However, higher temperature makes the loss more tolerant to different samples having similar embeddings and thus leads to less discriminative embeddings [21], which may cause poor extraction performance in TSE. We thus focus here on investigating methods to effectively use audio-only data for text-queried TSE without re-training the existing CLAP models.

Although the modality gap exists, relatively low positive cosine similarity (e.g., 0.4 on AudioCaps) suggests that the embeddings contain both some information shared between the two modalities and some domain-specific information. To prevent the TSE model from overfitting to such domain-specific features, we explore several methods that make the audio embeddings noisy, by either performing some data augmentation on the audio data or manipulating the audio embeddings.

Mixup: The audio embedding $\boldsymbol{q}_a$ is obtained from the mixture of the ground-truth audio $\boldsymbol{s}$ with another audio $\boldsymbol{s}'$, instead of from $\boldsymbol{s}$. Mixup was shown to be effective in AudioLDM, where a text-queried audio generation model is trained using only audio data [22]. We uniformly sample a signal-to-noise ratio (SNR) between the two audio signals from -5 to +5 dB.

SpecAugment: Some time-frequency (TF) bins of the mel spectrogram input to the CLAP audio encoder are zeroed out. Frame and frequency masking are done twice, where the number of frames or frequency bins are chosen from [0, 64] or [0, 2], respectively.

PCA: Audio embeddings $\boldsymbol{q}_a$ are projected into a d-dimensional space ($d < D$), where the projection matrix is trained with a small amount of text embeddings. We use 1000 text embeddings from the AudioCaps dataset to obtain the projection matrix, and set $d = 16$.

PCA-inv: Inverse-PCA is applied after PCA. Unlike normal PCA, the query vector is D-dimensional.

Gaussian noise: Scaled Gaussian noise is added to the audio embeddings. It has been shown that this method is effective for alleviating the gap between image and text embeddings in CLIP [23] or that between text and audio in CLAP [24]. Unlike [23], we regard the scale α as a hyper-parameter and manipulate embeddings as $\boldsymbol{q}_a \leftarrow \boldsymbol{q} + \alpha\boldsymbol{v}/||\boldsymbol{v}||$, where $\boldsymbol{v}$ is zero-mean Gaussian noise. Interestingly, we found that this method improves the robustness of the TSE model even in the text-queried training. Based on preliminary experiments, in each forward pass, we randomly choose α from [1.5, 3.5] when using audio embeddings and [1.0, 2.0] when using text embeddings.

Dropout: p-percent of the dimensions of the CLAP embeddings are dropped out (zeroed out) by Bernoulli dropout. The goal is to prevent the extraction model from overfitting to the domain-specific features by removing some information randomly. Based on preliminary experiments, we randomly choose p from [75.0, 95.0] when using audio embeddings and [25.0, 75.0] when using text embeddings.

III. Related work

This work is inspired by CLIPSep [14], where image-audio pairs without captions are leveraged for training text-queried TSE models by utilizing the CLIP encoder. CLIPSep also suffers from the modality gap in CLIP and fails to train conditional TSE models. Instead, CLIPSep trains a TSE system which consists of an unconditional separation module and a conditional post-mixing module. In contrast, our approach is easily applicable to conditional TSE models. In addition, while complicated training pipeline to estimate out-of-screen sounds is necessary when using an image as a query, we show that audio-queried training with a very simple modification is as effective as text-queried training.

Several prior works on text-queried TSE use CLAP as query encoder. In [11] and [25], both text and audio embeddings are used to improve performance or accept audio queries during inference. While these works assume that they have a text-audio pair for each data, our goal is to leverage audio-only data. Closest to our work, [12] tried audio-only training using the CLAP encoder but the training was not successful[1] In contrast, we show that audio-only data can be effectively used with simple embedding manipulation methods.

[1] Note that some results in [12] contradict those in [11] and ours. This was likely caused by an issue (now fixed) in the associated code when extracting audio embeddings.

For the text-queried audio generation task, AudioLDM [22] with a CLAP encoder was successfully trained using audio-only data. The modality gap is alleviated by the mixup augmentation. In the image field, [23] proposed to train an image captioning model using only text data, where the encoder is a pre-trained CLIP and the decoder is learnable. The modality gap is successfully avoided by injecting Gaussian noise into the CLIP text embedding. In a similar way, text-only training of an audio captioning model using a pre-trained CLAP has been achieved in [24]. Inspired by these works, we investigate methods to leverage audio-only data in text-queried TSE, which has not been achieved so far.

IV. Experiments

A. Datasets

We use the following two datasets for training. During training, input mixtures are created by uniformly sampling two audio signals and mixing them on the fly, where the signal-to-noise ratio is randomly chosen from -5 to +5 dB. All the signals are resampled to 32 kHz, following [12].

AudioCaps [20] includes 10-second audio clips from AudioSet and their human-annotated natural language captions. AudioCaps was used for the CLAP training [18], [19]. Following the original AudioCaps split, we use 49,827 and 495 audio clips for training and validation, respectively.

VGGSound [26] includes pairs of single-class audio clips and their class labels. Unlike AudioCaps, VGGSound was not used for CLAP training. However, it shares audio clips with AudioSet [27], which was used to train CLAP, so we removed the shared clips. Thereafter, all mentions of VGGSound refer to the set without the AudioSet clips. After the filtering, we split the original training data of VGGSound into 169,221 training and 2,500 validation clips. Since VGGSound does not have natural language captions, we use "*this is the sound of* $\{$*class*$\}$" when using text queries.

For testing, we use 6 datasets introduced in [28]: AudioCaps [20], Clotho v2 [29], VGGSound [26], AudioSet [27], ESC50 [30], and MUSIC [31]. Each mixture contains two sources from each dataset. AudioCaps and Clotho v2 have natural language captions, while the other four only have class labels and "*this is the sound of* $\{$*class*$\}$" is used as the query. MUSIC contains instrumental sounds, while the others are mainly composed of environmental sounds. Please refer to [12] and [28] for more details.

B. Models

As an extraction model, we test two backbones. Both operate in the short-time Fourier transform (STFT) domain, where a Hann window with length an L_w ms and hop size L_h ms is used.

Conformer [32] is the main backbone we use in our investigation as it is reported to work well on environmental sound separation [33] and is light-weight compared with the other model we consider. The model receives a magnitude spectrogram as input and estimates a real-valued TF mask for the target source. We use the mixture phase for resynthesis. It has 16 Conformer encoder layers with 4 attention heads, an attention hidden size of 256, and a feed-forward-network hidden size of 1024. A conditioning block, composed of a linear layer, Swish activation [34], a FiLM layer [35], and another linear layer, is placed before each Conformer block to incorporate the conditioning information. For the STFT, we set $L_w = 32$ and $L_h = 16$.

ResUnet is a state-of-the-art (SoTA) backbone in text-queried TSE. We use the same model as in [12], where FiLM is used as the conditioning layer. Receiving a complex spectrogram as input, the model estimates a magnitude mask and a phase residual component for the target source, where $L_w = 64$ and $L_h = 10$. Please refer to [12] for more details.

As query encoders, we use LAION-CLAP [18][2]. It was pre-trained on a large-scale audio-text dataset and its text encoder is often employed in text-queried TSE [12], [25]. To examine the effectiveness of audio-queried training on multiple CLAP models, we also test Microsoft-CLAP (MS-CLAP) [19][3].

C. Training/evaluation details

We train the model for around 400k training steps. We use the AdamW optimizer [36] with a weight decay factor of 1e-2. The learning rate is linearly increased from 0 to 2e-4 in the Conformer and 1e-3 in the ResUnet for the first 4k steps, kept constant for 160k steps, and then decayed by 0.9 every 10k steps. Gradient clipping is applied with a maximum gradient L_2-norm of 5. The batch size is 32 and the input mixture is 10 s long. As the loss function, we use the negative scale-invariant signal-to-distortion ratio (SI-SDR) [37].

D. Main results

Table I shows the evaluation results of the Conformer+LAION-CLAP model trained on AudioCaps (`A*`), VGGSound (`V*`), or both datasets (`AV*`). Again, note that AudioCaps has natural language caption and is used for the CLAP training, while VGGSound only has class labels and is not used for the CLAP training. In `A10`, `A11`, `V7`, and `V8`, we randomly choose either text or audio query in each training step.

First, compared with the text-queried training (`A0`), the audio-queried training (`A3`) gives much worse performance due to the modality gap in the CLAP encoder (see Section II-B). However, using some data augmentation or embedding manipulation methods can alleviate the problem (`A4–A9`). Simple embedding manipulations, namely Gaussian noise and dropout, performed the best among options, and audio-queried training with these manipulations achieves comparable or better performance than text-queried training (`A0` vs. `A8,A9`). It is also worth noting that these methods improve the performance of text-queried training on out-of-domain data (`A0` vs. `A1,A2`). We believe this is because they augment the text embeddings by adding or removing some noise and have a similar effect to caption augmentation [13], which makes the TSE model more robust against a variety of captions. Comparing `A0` and `A10`, we observe that using both text and audio as query during training improves the performance over text-only training, which is in line with the results in [11]. We again confirm the performance gain when using dropout (`A10` vs. `A11`), but using both text and audio queries does not improve performance over audio-only training (`A10` vs. `A9`).

We observe a similar trend when training on the VGGSound dataset: Gaussian noise or dropout on CLAP embeddings is effective in all cases (when using text, when using audio, and when using both). Comparing `A0–A2` and `V0–V2`, we observe that models work best on test data whose caption style (natural language or "*this is the sound of* $\{$*class*$\}$") matches that used in training. Interestingly, the overfitting to the caption style can be mitigated by using audio queries during training, and overall performance gets better. This result suggests that using audio as query during training can be beneficial for text-queried TSE even when we have class labels.

Finally, we consider the case where we have a certain amount of text-audio pairs (e.g., AudioCaps) and different audio-only data (e.g., VGGSound). Results in `A0` and `AV4` demonstrate that additional

[2] https://huggingface.co/lukewys/laion_clap/blob/main/music_speech_audioset_epoch_15_esc_89.98.pt

[3] https://huggingface.co/microsoft/msclap/blob/main/CLAP_weights_2023.pth

TABLE I
SI-SDR [dB] OF CONFORMER+LAION-CLAP MODEL ON TEST SETS.

			Captions		"This is the sound of {*class*}"			
ID	Method	Training data	AudioCaps	Clotho	VGGSound	AudioSet	ESC50	MUSIC
A0	-	AC-Text	7.6	5.1	6.2	2.3	8.3	0.8
A1	Gaussian noise	AC-Text	7.7	5.7	6.8	2.9	9.4	–0.6
A2	Dropout	AC-Text	7.9	5.8	6.9	2.9	9.2	0.2
A3	-	AC-Audio	5.1	1.7	0.8	–0.5	3.6	–1.2
A4	Mixup	AC-Audio	6.3	3.2	4.5	1.6	6.6	3.0
A5	SpecAug	AC-Audio	4.8	3.0	–0.3	–0.5	3.0	–0.3
A6	PCA	AC-Audio	6.8	2.8	4.7	0.4	5.5	0.1
A7	PCA-inv	AC-Audio	6.9	3.8	4.8	0.6	6.7	–0.5
A8	Gaussian noise	AC-Audio	**8.1**	6.3	6.9	3.4	9.3	1.2
A9	Dropout	AC-Audio	**8.1**	6.4	7.0	3.5	9.4	2.0
A10	-	AC-Text-Audio	7.9	6.0	6.5	3.2	9.1	1.3
A11	Dropout	AC-Text-Audio	8.0	6.4	7.0	3.5	9.6	0.7
V0	-	VGG-Text	5.0	4.0	8.6	3.9	9.5	9.4
V1	Gaussian noise	VGG-Text	5.8	5.1	8.5	4.5	9.7	8.8
V2	Dropout	VGG-Text	6.1	4.7	**8.8**	4.5	9.9	9.2
V3	-	VGG-Audio	4.2	3.0	2.3	1.3	4.0	3.5
V4	Mixup	VGG-Audio	6.2	4.3	6.1	2.9	8.0	8.3
V5	Gaussian noise	VGG-Audio	7.4	6.4	8.4	4.5	9.8	8.6
V6	Dropout	VGG-Audio	7.5	6.5	8.5	4.5	10.0	8.7
V7	-	VGG-Text-Audio	7.1	5.6	**8.8**	4.7	9.8	**9.6**
V8	Dropout	VGG-Text-Audio	7.2	6.3	8.5	**5.0**	9.9	9.0
AV0	-	AC-Text + VGG-Text	7.8	5.9	8.5	4.6	10.4	8.1
AV1	Dropout	AC-Text + VGG-Text	7.7	6.0	8.4	4.9	**10.2**	8.4
AV2	-	AC-Audio + VGG-Audio	5.4	2.9	2.3	1.4	4.7	6.3
AV3	Dropout	AC-Audio + VGG-Audio	7.8	6.5	8.4	4.7	10.0	7.9
AV4	-	AC-Text + VGG-Audio	7.9	5.9	7.3	4.2	9.5	7.8
AV5	Dropout	AC-Text + VGG-Audio	7.8	**6.6**	8.5	4.8	10.1	8.1

TABLE II
SI-SDR [dB] ON TEST SETS WHEN TRAINING A **RESUNET**+LAION-CLAP MODEL ON AUDIOCAPS. [4]

		Captions		"This is the sound of {*class*}"			
Query	Dropout	AudioCaps	Clotho	VGGSound	AudioSet	ESC50	MUSIC
Text		6.3	3.1	4.2	1.5	5.5	–0.4
Text	✓	6.9	4.2	6.5	3.2	8.4	0.3
Audio		2.9	1.1	1.1	0.5	2.5	**0.8**
Audio	✓	**7.1**	**4.9**	**7.2**	**3.9**	**9.1**	0.6
AudioSep [12]		7.2	5.2	9.0	6.9	8.8	9.4

audio-only data helps even without dropout. Still, dropout contributes to the performance gain (AV4 vs. AV5) and makes audio-queried training work as well as text-queried training (AV5 vs. AV0). This result suggests that we can effectively incorporate audio-only data to improve text-queried TSE systems.

E. Ablation study

We also trained ResUnet+LAION-CLAP and Conformer+MS-CLAP models to see if we observe a similar trend in other models. Based on the results of Table I, we use dropout as embedding manipulation.

Table II shows the evaluation results of ResUnet+LAION-CLAP trained on the AudioCaps data. The results demonstrate that dropout is also effective on the ResUnet extractor, which implies that audio-queried training with dropout is likely to be effective regardless of the architecture. We also list the performance of AudioSep [12] since we use the same model architecture[5]. Although AudioSep is trained on a much larger dataset, the audio-queried training achieves comparable performance on several test sets, which suggests that we may observe further performance gain over AudioSep by using the same dataset and audio-queried training.

TABLE III
SI-SDR [dB] ON TEST SETS WHEN TRAINING A CONFORMER+**MS-CLAP** MODEL ON AUDIOCAPS.

		Captions		"This is the sound of {*class*}"			
Query	Dropout	AudioCaps	Clotho	VGGSound	AudioSet	ESC50	MUSIC
Text		7.5	5.2	6.6	2.2	8.4	–0.2
Text	✓	**7.7**	5.7	7.1	2.6	9.0	2.0
Audio		3.1	1.6	3.8	0.5	5.9	2.1
Audio	✓	7.4	**6.4**	**7.8**	**3.1**	**9.9**	**3.3**

Table III shows the evaluation results of Conformer+MS-CLAP trained on the AudioCaps data. Again, audio-queried training with dropout gives the best performance. Although the training data and the encoder architecture of MS-CLAP are different from LAION-CLAP, the results show that MS-CLAP also has a modality gap problem and dropout helps to alleviate it.

V. CONCLUSION

We investigated methods allowing us to incorporate audio-only data for training text-queried TSE models. Although the CLAP models often have a modality gap and the TSE models easily overfit to audio queries, simple embedding manipulation methods such as dropout greatly alleviate the problem. Through experiments using multiple TSE models, we demonstrated that audio-queried training with dropout is as effective as text-queried training. In future work, we plan to scale up the training data by leveraging large-scale in-the-wild audio-only data.

[4] AudioSep uses the same architecture but trained on a much larger dataset.

[5] For fair comparison, all test scores are effectively computed on the same test data as the AudioSep official repository [28]. However, we do note that the reproduced scores for AudioSep do not match the original paper [12].

References

[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[2] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020.

[3] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021.

[4] K. Saijo, G. Wichern, F. G. Germain, Z. Pan, and J. Le Roux, "TF-Locoformer: Transformer with local modeling by convolution for speech separation and enhancement," *arXiv preprint arXiv:2408.03440*, 2024.

[5] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[6] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[7] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

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

[9] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. Interspeech*, 2020, pp. 1441–1445.

[10] 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, pp. 1801–1805.

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

[12] 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.

[13] D. H. Lee, Y. Song, and H. K. Kim, "Performance improvement of language-queried audio source separation based on caption augmentation from large language models for DCASE Challenge 2024 Task 9," *arXiv preprint arXiv:2406.11248*, 2024.

[14] H.-W. Dong, N. Takahashi, Y. Mitsufuji, J. McAuley, and T. Berg-Kirkpatrick, "CLIPSep: Learning text-queried sound separation with noisy unlabeled videos," in *Proc. ICLR*, 2023.

[15] 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 *Proc. ICML*, 2021, pp. 8748–8763.

[16] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. ICASSP*, 2023, pp. 1–5.

[17] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Proc. NeurIPS*, vol. 35, 2022, pp. 17 612–17 625.

[18] 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 *Proc. ICASSP*, 2023, pp. 1–5.

[19] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. ICASSP*, 2024, pp. 336–340.

[20] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. NAACL*, 2019, pp. 119–132.

[21] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proc. CVPR*, 2021, pp. 2495–2504.

[22] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: text-to-audio generation with latent diffusion models," in *Proc. ICML*, 2023, pp. 21 450–21 474.

[23] D. Nukrai, R. Mokady, and A. Globerson, "Text-only training for image captioning using noise-injected CLIP," *arXiv preprint arXiv:2211.00575*, 2022.

[24] S. Deshmukh, B. Elizalde, D. Emmanouilidou, B. Raj, R. Singh, and H. Wang, "Training audio captioning models without audio," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 371–375.

[25] 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.

[26] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A large-scale audio-visual dataset," in *Proc. ICASSP*, 2020, pp. 721–725.

[27] 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 *Proc. ICASSP*, 2017, pp. 776–780.

[28] 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," 2023. [Online]. Available: https://github.com/Audio-AGI/AudioSep

[29] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. ICASSP*, 2020, pp. 736–740.

[30] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Multimedia*, 2015, pp. 1015–1018.

[31] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. ECCV*, 2018, pp. 570–586.

[32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[33] K. Saijo and T. Ogawa, "Remixing-based unsupervised source separation from scratch," in *Proc. Interspeech*, 2023, pp. 1678–1682.

[34] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[35] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, vol. 32, no. 1, 2018.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2018.

[37] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.