# LANGUAGE-QUERIED AUDIO SOURCE SEPARATION ENHANCED BY EXPANDED LANGUAGE-AUDIO CONTRASTIVE LOSS

*Hae Chun Chung*

KT Corporation
Acoustic Processing Project, AI Tech Lab
Seoul, Republic of Korea
hc.chung@kt.com

*Jae Hoon Jung*

KT Corporation
Acoustic Processing Project, AI Tech Lab
Seoul, Republic of Korea
hoony.jung@kt.com

## ABSTRACT

Audio sources recorded for specific purposes often contain extraneous sounds that deviate from the intended goal. Re-recording to achieve the desired result is expensive. However, separating the target source from the original audio source based on natural language queries would be much more efficient. However, audio source separation with natural language queries is a complex task. To address this, the DCASE 2024 Challenge Task 9 proposed language-queried audio source separation (LASS). This paper aims to tackle LASS by proposing an extended language-audio contrastive learning approach. To align the separated output audio with the target text and target audio, we first designed audio-to-text contrastive loss and audio-to-audio contrastive loss, respectively. By leveraging the characteristics of contrastive learning, we combined these two losses into an extended audio-to-multi contrastive loss. Our model, trained with this loss, improves the signal-to-distortion ratio (SDR) by more than 30% compared to the baseline provided by the challenge.

***Index Terms***— Source Separation, Contrastive Learning

## 1. INTRODUCTION

In real-world scenarios, unintended and uncontrollable events frequently occur. During on-location content recording, numerous factors are managed to capture the desired purposes. Nevertheless, unwanted elements often contaminate the recorded audio sources. Re-recording to achieve perfection is not only expensive but also challenging. If an AI model could separate the target source from the recorded audio source based on natural language queries, these costs could be significantly reduced. However, audio source separation with natural language queries is a complex task. Consequently, research in this field is limited, and existing performance levels are suboptimal [1, 2]. To address this, the DCASE 2024 Challenge Task 9 proposed language-queried audio source separation (LASS) [3]. This task focuses on developing a system that separates the target audio source from a mixed audio source based on a text description about the intended audio.

LASS-Net [1] first introduced the task of language-queried audio source separation (LASS), proposing an end-to-end neural network consisting of a text encoder, which takes a text description (target text) as input and outputs a text embedding, and a separator, which takes the mixed audio (a mixture of a target audio and a noise audio) and text embedding as inputs to predict the target audio. AudioSep [2] used a contrastive language-audio pre-training

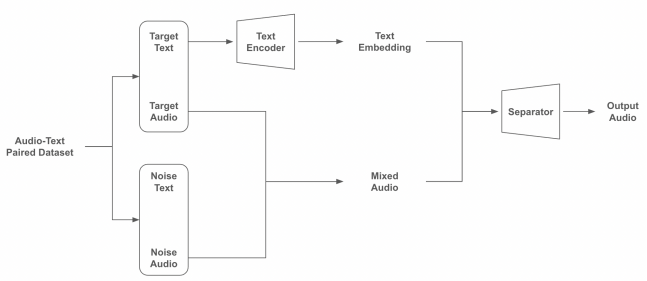


Figure 1: The figure above provides a schematic overview of our model. From the audio-text paired dataset, a pair consisting of target text and target audio, along with a pair of noise text and noise audio, are randomly sampled to ensure they do not overlap. The target audio and noise audio are mixed at a signal-to-noise ratio (SNR) ranging from -15 to 15 dB to create a mixed audio. The text encoder extracts a text embedding from the target text. The separator then takes this text embedding as a condition and the mixed audio, separating the output audio conditioned on the text embedding from the mixed audio.

model (CLAP) model [4] as the text encoder, which was frozen during training, and the separator was trained to predict phase residuals as well as a magnitude mask [5]. Furthermore, unlike LASS-Net, which was trained on a subset of the AudioCaps dataset [6], AudioSep was trained with large-scale audio datasets, leading to a significant performance improvement over LASS-Net. The baseline system for the DCASE 2024 Challenge Task 9 is based on the AudioSep model, but it only used the development set (Clotho [7] and augmented FSD50K [8] dataset) provided in the challenge for training data. This baseline system achieved a signal-to-distortion ratio (SDR) score of 5.708 when evaluated on the validation dataset provided in the challenge.

We also adopted a model structure consisting of a text encoder and a separator. The separator was same to ResUNet [5] setting used in AudioSep. For the text encoder, we used FLAN-T5 [9], an instruction-tuned large language model (LLM), instead of the CLAP model. FLAN-T5 was chosen based on its successful application as a text encoder in TANGO [10], which addresses the text-to-audio generation task. To train this system, we introduced three loss functions, and utilized a loss balancer [11] to stabilize the training. First, L1 loss was employed to align the separated audio wave-

form with the target audio waveform in the time domain. Second, to optimize performance in both the time and frequency domains, we utilized multi-scale mel-spectrogram loss [12, 13, 11, 14], applied across multiple time scales in the mel-spectrogram. Lastly, contrastive loss was introduced in addition to L1 loss and spectrogram loss.We designed three distinct contrastive losses using target audio, noise audio, target text, and noise text for output audio. To embed audio and text, CLAP model [4] was used. First, audio-to-text contrastive loss (A2T-CL) was introduced to increase the similarity between output audio and target text while reducing the similarity with other non-target texts within the mini-batch. The performance was further improved by combining audio-to-audio contrastive loss (A2A-CL), which applies to the target audio and other non-target audios within the mini-batch, with A2T-CL. Contrastive learning tends to improve performance as the comparison samples, especially negative samples, increases [15, 16]. For leveraging this, we designed the audio-to-multi contrastive loss (A2M-CL) by integrating A2A-CL and A2T-CL into a single expanded loss. A2M-CL encourages output audio to increase similarity for both the target text and the target audio while reducing similarity for other non-target texts and audios in the mini-batch. This doubles the number of comparison samples, both positive and negative samples, than A2A-CL or A2T-CL. We experimented for each method and achieved SDR scores of 7.030, 7.12, and 7.139, respectively. This is a performance improvement of more than 30% over the baseline model.

## 2. METHODS

### 2.1. Overview

Our system consists of two models: a text encoder and a separator. For the text encoder, we utilize FLAN-T5 [9], an enhanced version of the text-to-text transfer transformer (T5) model [17]. FLAN-T5 is initialized with a T5 checkpoint and fine-tuned with instructions and chain-of-thought reasoning, enabling it to extract robust text embeddings from text descriptions with its strong text representation capacity. TANGO [10], which tackles the text-to-audio generation task, demonstrated effectiveness of FLAN-T5 as the text encoder for cross-modal task.

The separator is the ResUNet model [18, 5], an advanced version of the UNet model. We used the same setting as ResUNet used in AudioSep [2]. The ResUNet model takes a mixed audio waveform and text embedding as input and separates the output audio waveform conditioned on the text from the mixed audio. The process begins with applying a short-time Fourier transform (STFT) to the waveform to extract the complex spectrogram, magnitude spectrogram and phase. The ResUNet model takes the complex spectrogram and outputs the magnitude mask and phase residual conditioned on the text embedding. The separated complex spectrogram is obtained by multiplying the STFT of the mixture with the predicted magnitude mask and phase residual. Finally, the separated complex spectrogram is converted back into an audio waveform using the inverse short-time Fourier transform (iSTFT).

### 2.2. Training Loss Terms

From the audio-text paired dataset, $N$ target pairs (target audio $d^{ta}$ and target text $d^{tt}$) and $N$ noise pairs (noise audio $d^{nt}$ and noise text $d^{nt}$ ) are randomly sampled to ensure they do not overlap. For creating mixed audio waveform $d^{ma}$, two audio waveforms are combined with a signal-to-noise ratio (SNR) ranging from -15 to 15 dB. The target text is forwarded into the text encoder to extract the text embedding. The separator then takes the mixed audio waveform and the text embedding, separating the output audio waveform $d^{oa}$ conditioned on the text from the mixture.

**L1 Loss** In the source separation task, it is crucial to extract the desired target sound source from a given mixture without altering its original characteristics. In other words, the closer the separated sound source is to the target sound source, the better the performance. To achieve this, minimizing the L1 distance between the target audio and separated audio over the time domain is commonly used due to its simplicity and effectiveness in universal source separation tasks. We also applied this approach. The equation is as follows:

$$\mathcal{L}_{time} = \left\| d^{ta} - d^{oa} \right\|_1 \tag{1}$$

**Spectrogram Loss** To optimize performance in both the time and frequency domains, we also employed a multi-scale mel-spectrogram loss [12, 13, 11, 14] applied across multi time scales in the mel-spectrogram. This loss is calculated based on the distance in the mel-spectrogram, which is derived from the short-time Fourier transform (STFT) and converted to a mel scale that better captures human auditory characteristics. This approach enhances the perceptual quality of the output. Additionally, using loss functions on mel-spectrograms across multiple STFT scales enables the model to effectively capture the time-frequency distribution, significantly enhancing its overall performance.

$$\mathcal{L}_{freq} = \frac{1}{|\alpha| + |s|} \sum_{\alpha_i \in \alpha} \sum_{i \in e} \left\| \mathcal{S}_i(d^{ta}) - \mathcal{S}_i(d^{oa}) \right\|_1 + \alpha_i \left\| \log \mathcal{S}_i(d^{ta}) - \log \mathcal{S}_i(d^{oa}) \right\|_2 \tag{2}$$

where $\mathcal{S}_i$ is a 64-bins mel-spectrogram using a normalized STFT with window size of $2^i$ and hop length of $2^{i-1}$, $e = 6, ..., 12$ is the set of scales, and $\alpha$ represents the set of scalar coefficients balancing between the L1 and L2 terms, $\alpha_i = \sqrt{2^{i-1}}$. Here, $|\alpha|$ denotes the sum of the elements of the $\alpha$ set, and $|s|$ is the number of scales.

**Audio-to-Text Contrastive Loss** The output audio of the text-conditioned audio source separation should match the target audio, for which L1 loss and spectrogram loss were used. Additionally, the output audio must involve all the content of the target text while excluding any content not present in the target text. To achieve this, we implemented an audio-to-text contrastive loss (A2T-CL) using the contrastive language-audio pre-training (CLAP) model [4]. CLAP was trained to align audio and text by projecting them into a shared feature space. Firstly, we designed the loss so that the output audio attracts its corresponding target text as positive and repels other target texts within the mini-batch as negative in the shared feature space of CLAP model. Contrastive learning tends to improve performance as the comparison samples, especially negative samples, increases [15, 16]. To leverage this, we additionally use noisy texts within the mini-batch as negative examples. This approach encourages the output audio to be distinguishable from various other texts while accurately fitting the target text. The equation is as follows:

$$\mathcal{L}_{a2t} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(f_i^{oa} \cdot f_i^{tt} / \tau)}{\sum_{k=1}^{N} \left\{ \exp(f_i^{oa} \cdot f_k^{tt} / \tau) + \exp(f_i^{oa} \cdot f_k^{nt} / \tau) \right\}} \tag{3}$$

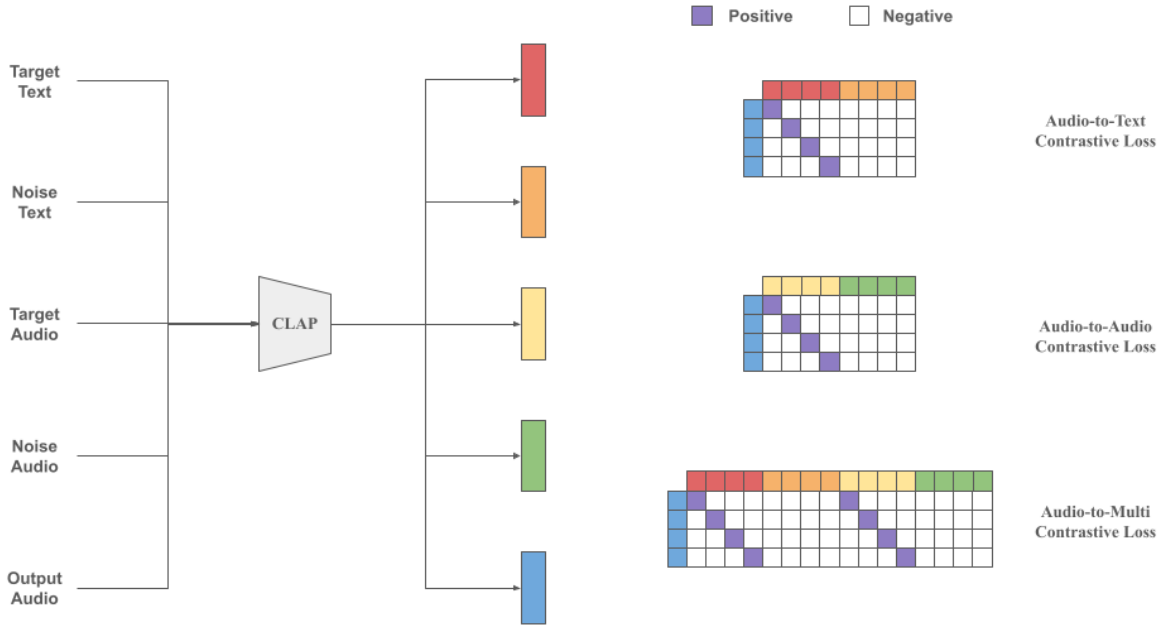


Figure 2: Each rectangle in red, orange, yellow, green, and blue represents the features of target texts, noise texts, target audios, noise audios, and the output audios of the separator, all embedded using the CLAP model. The matrices on the right schematically illustrate three types of contrastive loss with a mini-batch size of 4. In these matrices, purple spaces indicate positive relationships, while white spaces indicate negative relationships. The output audio has positive relationships with its corresponding target text and target audio, whereas all other texts and audios within the mini-batch are considered negative relationships.

where $f^{oa}$ is a feature with output audio embedded using audio encoder of CLAP model, and $f^{tt}$ and $f^{nt}$ are features with target text and noise text embedded using text encoder of CLAP model. And $\tau$ is a scalar temperature parameter.

**Audio-to-Audio Contrastive Loss** The concept of A2T-CL, which encourages output audio to contain only the content of the target text, can also be applied to audios (target audios and noise audios). Therefore, it is possible to design an audio-to-audio contrastive loss (A2T-CL) using these.

$$\mathcal{L}_{a2a} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(f_i^{oa}\cdot f_i^{ta}/\tau)}{\sum_{k=1}^{N}\left\{\exp(f_i^{oa}\cdot f_k^{ta}/\tau)+\exp(f_i^{oa}\cdot f_k^{na}/\tau)\right\}} \tag{4}$$

where $f^{ta}$ and $f^{na}$ are features with target audio and noise audio embedded using audio encoder of CLAP model.

**Audio-to-Multi Contrastive Loss** As aforementioned, contrastive learning shows better performance as the number of the comparison samples increases. To take advantage of this, we integrated audio-to-text contrastive loss and audio-to-audio contrastive loss into a single expanded loss: audio-to-multi contrastive loss (A2M-CL), effectively doubling the number of the comparison samples. This causes the output audio to pull closer to its corresponding target text and target audio while pushing away from all remaining target texts, noise texts, target audios, and noise audios within the mini-batch. As a result, the output audio maximizes its similarity to both the target text and target audio.

$$a2t_i = \sum_{k=1}^{N}\left\{\exp(f_i^{oa}\cdot f_k^{tt}/\tau)+\exp(f_i^{oa}\cdot f_k^{nt}/\tau)\right\} \tag{5}$$

$$a2a_i = \sum_{k=1}^{N}\left\{\exp(f_i^{oa}\cdot f_k^{ta}/\tau)+\exp(f_i^{oa}\cdot f_k^{na}/\tau)\right\} \tag{6}$$

$$\mathcal{L}_{a2m} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left\{\log\frac{\exp(f_i^{oa}\cdot f_i^{tt}/\tau)}{a2t_i+a2a_i}+\log\frac{\exp(f_i^{oa}\cdot f_i^{ta}/\tau)}{a2t_i+a2a_i}\right\} \tag{7}$$

**Loss Balancer** Encodec [11] introduced a loss balancer to stabilize the training by adjusting the loss weights based on various scales of gradients from the model. We used a loss balancer to stabilize the model training with various losses. The gradient $\frac{\partial l_i}{\partial d^{oa}}$ of the loss based on the output $d^{oa}$ is recalculated using the following equation, incorporating the weights $\lambda_i$ for the loss and reference norm $R$.

$$\tilde{g}_i = R\frac{\lambda_i}{\sum_j \lambda_j}\cdot\frac{g_i}{\left\langle \|g_i\|_2 \right\rangle_\beta} \tag{8}$$

where $\left\langle \|g_i\|_2 \right\rangle_\beta$ is the exponential moving average of $g_i$. We take $R = 1$ and $\beta = 0.999$. All the model losses fit into the balancer. The model is then backpropagated to $\sum_i \tilde{g}_i$ instead of the original $\sum_i \lambda_i g_i$.

### 2.3. Proposed Systems

We propose a total of three systems. The process by which data is preprocessed and fed forward to the model in all systems is the same as mentioned in Section 2.1. The primary difference between each system lies in the configuration of losses during the training process, particularly the type of contrastive loss. The configuration of the losses for each system in our training was defined as follows. All weights $\lambda$ for the losses are set 1.

$$System_1 = \lambda_1 \mathcal{L}_{time} + \lambda_2 \mathcal{L}_{freq} + \lambda_3 \mathcal{L}_{a2t} \quad (9)$$

$$System_2 = \lambda_1 \mathcal{L}_{time} + \lambda_2 \mathcal{L}_{freq} + \lambda_3 \mathcal{L}_{a2t} + \lambda_4 \mathcal{L}_{a2a} \quad (10)$$

$$System_3 = \lambda_1 \mathcal{L}_{time} + \lambda_2 \mathcal{L}_{freq} + \lambda_3 \mathcal{L}_{a2m} \quad (11)$$

## 3. SETTING

### 3.1. Training Data

A total of four datasets were used for model training: AudioCaps [6], WavCaps [19], Clotho v2 [7], and FSD50K [8]. For the WavCaps dataset, only data belonging to AudioSet were used. The combined dataset comprises a total of 216,398 audio clips, amounting to approximately 580 hours. The following procedure was employed to generate mixed audio:

1. Random Selection: Target and noise audio clips were randomly selected to ensure no overlap within the entire dataset.
2. Mono Conversion: If an audio clip had 2 channels, the average of the two channels was calculated to convert it into a mono clip.
3. Resampling: Audio clips with a sampling rate different from 16 kHz were resampled to 16 kHz.
4. Length Adjustment: If an audio clip exceeded 10 seconds in length, it was randomly truncated to 10 seconds. If it was shorter than 10 seconds, zero padding was added to the end to make it 10 seconds long.
5. Mixing: The pre-processed target audio clip and a noise audio clip were mixed with signal-to-noise ratios (SNR) ranging from -15 dB to 15 dB to produce a mixed audio clip.

### 3.2. Model

The text encoder for embedding the text is used pre-trained FLAN-T5 model [9], and all parameters were frozen. AdamW optimizer [20] with a learning rate of 0.0003 is used for training the separator with the batch size of 25. $\tau$ was all set to 0.1 for the contrastive loss.

### 3.3. Test Data

To evaluate the performance of the model, validation (synth) dataset provided in DCASE2024 Challenge Task9 [3] was used.

### 3.4. Metric

To compare the performance of language-queried audio source separation (LASS), we used three objective metrics that are commonly used in the field of source separation: signal-to-distortion ratio (SDR), signal-to-distortion ratio enhancement (SDRi), and scale-invariant SDR (SI-SDR) [21].

## 4. RESULTS

The language-queried audio source separation (LASS) task is a nascent field with limited prior research. However, due to its high usability and future potential, the DCASE Challenge adopted this task as Task 9 for this year. We participated in Task 9 of the DCASE 2024 Challenge to officially demonstrate the performance of our model, specifically designed for LASS

| | SDR | SDRi | SI-SDR |
|---|---|---|---|
| Baseline | 5.708 | 5.673 | 3.862 |
| System1 | 7.030 | 6.995 | 5.368 |
| System2 | 7.124 | 7.089 | 5.593 |
| System3 | 7.139 | 7.104 | 5.504 |

Table 1: The comparison of baseline model and our proposed model on validation set.

We compared our system with the baseline model provided by challenge. The baseline provided in the challenge was based on the AudioSep model. Table 1 shows the performance comparison between the baseline model and our proposed systems using the challenge validation set. Our proposed systems show significant performance improvements across all three metrics. While the baseline provided for the challenge achieved a signal-to-distortion ratio (SDR) score of 5.708, our systems achieved SDR scores of 7.030, 7.124, and 7.139, respectively. This represents a remarkable performance improvement of over 30% compared to the baseline. In language-queried audio source separation (LASS), it is crucial to precisely match the output audio to the target audio. Additionally, we demonstrate that aligning the output audio more closely with both the target text and target audio in the feature space using contrastive learning enhances performance. We also show the effectiveness of the audio-to-multi contrastive loss, which leverages the characteristics of contrastive learning by integrating audio-to-text and audio-to-audio contrastive losses. This approach leverages the advantage of having more negatives, significantly improving the model's effectiveness.

| | SDR | SDRi | SI-SDR |
|---|---|---|---|
| Baseline | 5.799 | 5.693 | 3.873 |
| System1 | 7.302 | 7.195 | 5.628 |
| System2 | 7.186 | 7.080 | 5.526 |
| System3 | 7.118 | 7.012 | 5.301 |

Table 2: The comparison of baseline model and our proposed model on evaluation set.

We received a score by submitting the results of each system on the evaluation set to the challenge. Contrary to expectations, the evaluation in the evaluation set came out opposite to the evaluation in the validation set. We doubt whether this is an overfitting on the validation set. We leave it as a future work. However, nevertheless, the systems we proposed showed significant performance improvement compared to the baseline. The performance can be improved using the contrastive learning method we designed simply without modification to the model architecture.

## 5. REFERENCES


[1] 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," *arXiv preprint arXiv:2203.15147*, 2022.

[2] 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.

[3] https://dcase.community/challenge2024/task-language-queried-audio-source-separation.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[10] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[11] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[12] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[13] A. Gritsenko, T. Salimans, R. van den Berg, J. Snoek, and N. Kalchbrenner, "A spectral energy distance for parallel speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 062–13 072, 2020.

[14] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[18] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation," *arXiv preprint arXiv:2109.05418*, 2021.

[19] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[21] 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.