TOWARDS DISTANCE-AWARE SYNTHETIC AUDIO MIXTURES FOR UNIVERSAL SOUND SEPARATION

Wonjun Park, Tuan M. Dang, Kenny Q. Zhu

Computer Science and Engineering
University of Texas at Arlington
Arlington, TX, USA

ABSTRACT

A machine learning model performs a designated task based on the training data. Albeit existing universal sound separation models which heavily rely on a mix-and-separate framework due to a lack of training data show a feasible performance to some extent, the performance is still limited to use in strict practical applications. With a hypothesis that the gap is emerged by different distributions between a random and a real-world, this paper introduces an audio mixing strategy on audio-text pair datasets. A distance knowledge between audio sources is derived from Large Language Models, determining how far apart they should be in a mixed output. Pairs of mixtures and their corresponding ground truths are created as a result. To validate our hypothesis and the effectiveness of the strategy, human evaluation is mainly conducted on pure recordings manually curated from multiple datasets. A result of the synthesized benchmark randomly created by a previous work is also reported to discuss the impact of the naturalistic audio in the sound source separation field.

***Index Terms*—** Data Synthesis, Universal Sound Separation, Large Language Models, Knowledge Transfer

1. INTRODUCTION

Universal Sound Separation (USS) [1] is a task to disentagle individual sound sources from a given single-channel recording that may contain an arbitrary and unknown set of overlapping sound events. Practical demand for the USS task is obvious across multiple disciplines. Animal Language Processing [2, 3, 4], for example, requires to separate a target animal vocal from various sound sources. Recordings are collected from both the wild and the Internet, containing unwanted sound sources (e.g., *vocalizations from other species*, *wind noise*, and *human speech*) as well. All recordings are precious for the researchers to analyze animal vocals to imply the several situations as much as possible, since languages are susceptible to the context and the environment.

After Mixture Invariant Training (MixIT) [5] showed the potential of Mixture of Mixtures (MoMs) as a training data in USS, not only leveraging unannotated audios from the Internet but also eliminating the need of the ground truth audio, synthetic data becomes crucial especially in this field, where the accessibility to the ground truth audio is virtually impossible or too expensive. Many previous studies have utilized MoMs as synthetic input for their separation models.

To our knowledge, all existing studies typically generate MoMs via random mixing, without considering whether the resulting acoustic signal is realistic or not. Even though this mix-and-separate strategy has been achieving a performance to some extent, the resulting MoMs is often unrealistic and unnatural. Since data strongly shapes what a model learns, an indiscriminate mixing strategy risks biasing the network toward implausible source combinations, a potential obstacle in practice. Take two mixtures, *Whale vocalization* and *Vehicle horn, car horn, honking* from the labeled audios on AudioSet [6], for example. If they are mixed together with the existing random strategy, the model is led to learn such unrealistic data, although it is nearly impossible that those two sound events coincide in a natural environment.

This paper aims to create a more realistic MoMs by leveraging distance knowledge regarding sound events. We believe that such realistic mixtures ultimately lead separation models to achieve more robust performance in real-worlds scenarios. Recently, Large Language Model (LLM) is acknowledged to have a comprehensive bank of world knowledge by reading the most existent texts, an enormous amount that a human being can not read during their entire lifetime. Hinged on the capacity of the LLM, we ask the LLM to answer the following question with labels or captions from audio-text pair datasets; To what distance does a sound source naturally arise alongside the given sound source? The answer guides to better and more plausible sound mixtures.

Human evaluation is crucially conducted for comparison, supporting our approach as being indeed effective in real mixtures. Besides, we also report a synthesized benchmark result created by previous works, deriving some discussion about the effect of the naturalistic audio.

Our contributions are lied in the following three aspects:

- We propose a novel mixing framework to align training

This work is partially supported by NSF Award No. 2349713.

data hinged on distance knowledge (Sec. 2). It derives from LLM which has an abundant textual knowledge about the world.

- We show that a random distribution and a real-world distribution are different in sound source separation, a crucial insight especially for small datasets (Sec. 3).
- With the proposed strategy, we get better separation results over the previous random mixing strategy selected 2 times more at the most in human evaluation.

2. METHODOLOGY

Figure 1 gives an overview of the LLM mixing module for our distance-based mixing strategy. The module exploits LLM to determine how a sound source is dominant in a miuxture, amplifying its volume accordingly.

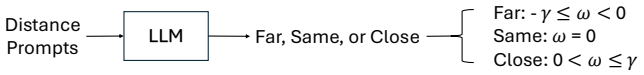

Fig. 1: LLM Mixing Module. The LLM determines a plausible distance with distance prompts.

2.1. Audio Synthesis with Distance Knowledge

Let's call a base audio a_i and a corresponding caption t_i among N audio-text pairs where $i \in N$. A candidate audio $a_{i^\star}$ is chosen among N, where $i^\star \neq i$. The relative loudness of the candidate audio against the base audio is determined by inferring the distance of the candidate sound source from the base audio source. The LLM is prompted with the caption t_i and the candidate caption $t_i^\star$ to answer a plausible distance among `far`, `same`, and `close`, a distance of $t_i^\star$ relative to t_i Each distance is mapped into each range of the loudness, randomly selecting a dB value ω, from each range; `far` $\rightarrow [-\gamma, 0)$ dB, `same` $\rightarrow 0$ dB, and `close` $\rightarrow (0, \gamma]$ dB, where γ is a hyperparameter that controls the range of ω. If the LLM answer is not `far`, `same`, or `close`, it feeds into the text encoder to compare the similarity against the three labels, guaranteeing one of the three answers is selected.

The dB value and the mixture derive the following value

$$\alpha = \sqrt{\frac{E_1}{E_2} \cdot 10^{\omega/10}}$$

$$m_i = a_i + \alpha \cdot a_{i^\star}$$

where α is a scaling factor to adjust the energy between the base audio a_i and the candidate audio $a_{i^\star}$, E_1 and E_2 are the energies of the base audio and the candidate audio, as well as m_i is the MoM of the i-th audio-text pair.

For example, if the LLM is received that a base audio clip is a *frog croaking* and a candidate audio is *rain falling*, the LLM is induced to answer `far` since the frog sound is mostly dominant during the rain. Note that we do not consider other acoustic factors such as frequency attenuation.

2.2. LLM Prompts

The following box presents the prompts in the LLM mixing module;

Distance Prompts

System Prompt
You will be given two captions:

1. A caption describing a *base* audio.
2. A caption describing a *candidate* audio.

Decide how loudly the candidate audio should be mixed with the base audio so the blend sounds the most natural.
Return **exactly one** of the following tokens and nothing else:

- far : candidate sounds distant / noticeably quieter than the base
- same : candidate is at roughly the same loudness as the base
- close : candidate sounds very near / noticeably louder than the base

User Prompt
USER: A base audio caption: relaxed music
A candidate audio caption: human speech
ASSISTANT: close
USER: A base audio caption: a child laughs
A candidate audio caption: a woman speaks
ASSISTANT: same
USER: A base audio caption: birds chirp
A candidate audio caption: a branch sways in the wild
ASSISTANT: far
USER: A base audio caption: t_i
A candidate audio caption: $t_i^\star$
ASSISTANT: *A plausible distance among far, same, or close*

Wrapping up the section, the mixing procedure in Sec. 2 are conducted for every mini-batch of training so that the model virtually sees different MoMs in every iteration.

3. EXPERIMENTS

We conduct comparisons on two major sound separation approaches across different model architectures: One is TDCN++ [1, 5], a mask generator in a waveform domain, and the other is ResUNet, a model that utilizes a U-Net ar-

chitecture in a spectrogram domain. MixIT, a unconditional sound separation algorithm, is used to train the TDCN++ model, and AudioSep [7], a text-guided conditional sound separation algorithm at a DCASE 2024 challenge, is used to train the ResUNet model. Note that MixIT is included to show that our approaches work regardless of the model architecture, although its strength is hinged on the availability of unannotated audio.

Experiments are conducted on a server with 2 NVIDIA RTX 4090 GPUs, each with 24 GB of memory, as well as two computational nodes installed with three and one NVIDIA A100 GPUs, respectively, each with 40 GB of memory.

3.1. Implementaion Details

The γ, a hyperparameter which indicates a range of dB to control a distance in Sec. 2.1, is selected as 15 in our implementation. Experiments reported in this paper use a publicly available LLM, Meta Llama 3.2 1B. Few-shot prompting is also utilized to guide the LLM not only to answer in a specific format but also to improve the quality of the answers. The details of the few-shot prompting are shown in Sec. 2.2. Text encoders, both the CS6 module and AudioSep, in our experiments are from the pretrained weights of Contrastive Language-Audio Pretraining (CLAP) [8] which aligns audio and text modalities.

All experiments are performed with a batch size of 18, and the training steps for each strategy in each model are kept the same to fairly compare the performance. The rest of the training details mostly follow the original paper of each model.

3.2. Datasets

3.2.1. Training Data

We download Clotho v2.1 [9], and FSD50K [10]. The length of the audio clips is varied from 0.3 second to 30 seconds. Training data of the separation models is composed by both the development set and eval set of FSD50K as well as the development, validation, and evaluation sets of Clotho. Every audio signal is resampled to 16 kHz and combined into a single channel before feeding into the separation models. AudioSep and MixIT are trained with 10-second and 5-second audios respectively, which randomly crops the audio clips from the training data while mixing.

3.2.2. Real-world Benchmark

Three different datasets are curated for human evaluation, containing 100 audio mixtures respectively;

Indoor+City is from a test set of AudioCaps [11]. Indoor+City aims to cover indoor and city sounds, such as human speech in a cafe, water running in a toilet, and a car honking in the street are included. **Outdoor+Wild** focuses on outdoor and wild sounds, collected from the AudioCaps test set as well. It accounts for outdoor and wild sounds like a bird chirping in the forest, a thunderstorm, and a river flowing. Last, **Music** is with a collection of 11 different sources of music from MUSIC [12], accommodated by a pair of two musical instruments such as a flute, a violin, and a trumpet. These enables more diverse and realistic mixtures than MUSDB18 [13].

3.2.3. Synthesized Benchmark

We report a synthesized benchmark result created by arbitrary mixing two sound events from ESC50 [14], consisted of 5-second audios with 50 distinguished sound events.

3.3. Results

3.3.1. Human Evaluation

We utilize the vote application (shown in Figure 2) to conduct human evaluation. The application prints the original caption of the audio which describes what sound sources emerge in the audio. The target caption is used in AudioSep to condition the separation model as well as the criteria of the separated result on both MixIT and AudioSep. In other words, a human evaluator, which we called a voter, determines which separated results is better based on the target caption.

Importantly, the candidates of the vote, *Random* versus *Distance*, are hidden from the voters and their sequence is randomly shuffled, printing *Option 1*, *Option 2*, and *Tie* buttons. The voters are asked to listen to the original audio and the separated results, and select the button below the UI. The spectrogram of each option is displayed to the voters as well to help them make a decision, if they encountered a situation where the audio is not enough to distinguish the separated results. Each vote adds one point to the chosen outcome, while a *Tie* vote contributes 0.5 points to each outcome.

We also calculate an agreement rate between two voters. There are three possible cases per clip: (1) One annotator votes *Random* and the other votes *Distance* which adds 0 to an agreement score, (2) One annotator prefers *Random* (or *Distance*) and the other votes Tie, which adds 0.5 to the score, and (3) both annotators vote for the same method, which adds 1 to the score. The agreement rate is then calculated as the total agreement score divided by the number of clips. The final agreement rate at Table 1 is averaged over all pairs of voters.

3.3.2. Automatic Evaluation

We adopt the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [15] metric as our primary evaluation criterion. The metric allows a quantifiably stable comparison between the signals than the Signal-to-Distortion Ratio (SDR). Table 1 shows both human and automatic evaluation results.

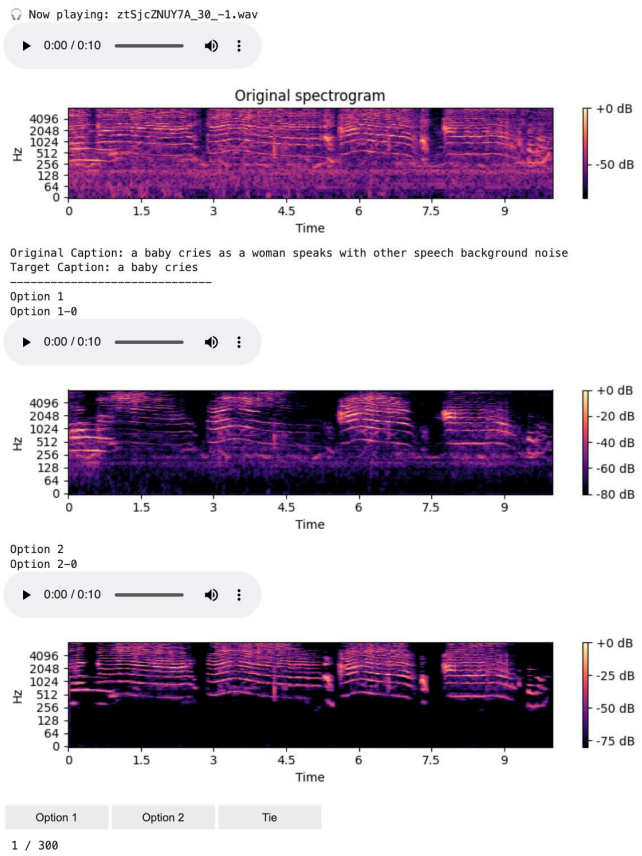

Fig. 2: UI of Vote Application for Human Evaluation. The first audio is the original audio. Option 1 and 2 are the hidden strategies between *Random* and *Distance*.

3.3.3. Discussion

Our distance-aware mixing introduces a world-knowledge prior that better matches real acoustic scenes. By aligning relative loudness with plausible source distances, the training distribution becomes more coherent, which smooths the loss landscape, accelerates convergence, and reduces attraction to spurious minima. The same realism, however, biases capacity toward high-probability real-world regions; performance can therefore degrade on synthetic tests built from unlikely, randomly paired, and uniformly loud sources. Given the ill-posed nature of sound separation, unconditional models are especially subject to such distribution shifts, reducing their hypothesis space. This is not a flaw of the approach but rather an expected and better bias for the intended domain, worse for an out-of-domain synthetic benchmark. In contrast, conditional models even benefit from improved compatibility between text semantics and acoustic context as well as the external knowledge. These findings imply that widely used randomly synthesized benchmarks undervalue methods that model real scenes, raising a new problem for the sound separation field.

	Benchmark	Random	Distance
AudioSep	Indoor+City	28.7%	**71.3%**
	Outdoor+Wild	25%	**75%**
	Music	38.3%	**61.7%**
	ESC50	1.597	**3.029**
MixIT	Indoor+City	43.3%	**56.7%**
	Outdoor+Wild	47.7%	**52.3%**
	Music	47%	**53%**
	ESC50	**8.292**	4.961
# Voters / Agreement Rate		4 / 69.1%	

Table 1: Human Preference and SI-SDR Results for AudioSep and MixIT; For each model, the first three rows represent human evaluations, and the fourth row is SI-SDR. The last row indicates the number of voters and their agreement rates integrated for both models.

4. RELATED WORK

Early deep learning approaches to single-channel source separation relied on supervised training with labeled and isolated audio sources, which severely limited the size and diversity of the training data. MixIT [5] broke this dependency with both its unsupervised training and MoMs, mixing two mixtures on-the-fly and training a network to separate a variable number of sources. However, the framework required 2^n computations as the number of sound sources increases and the order of the separated sound sources is not guaranteed so that it was not suitable to find a target sound source. On top of MixIT, AudioSep [7] extended a ResUNet architecturewith a text encoder to separate a target sound source from a mixture. Still, the performance of the model is so limited that the model is insufficient to be used in real-world scenarios.

HTDemucs [16] addressed how to mix sounds in music. With beat tracking and tempo estimation, they created a dataset for fine-tuning containing reasonable melodic pieces of music, leading the state-of-the-art performance on the MUSDB18 benchmark at that time. The heterogeneous target speech separation [17] utilized a conditional separation network, showing that features in a real-world setting, like the distance from the microphone or the language of the speaker, are useful for the separation task.

5. CONCLUSION

We present new synthesis strategy for sound source separation that leverage distance knowledge. With LLMs fulfilled with common-sense in a textual domain, we generate MoMs that closely resemble natural audio mixtures, enhancing the performance of separation models in real-world scenarios. Nevertheless, the LLM inference during training introduces a computational overhead. Future work will explore other possible knowledge that can be integrated into the mixing strategy, or reduce the computational cost.

6. REFERENCES

[1] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey, "Universal sound separation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 175–179.

[2] Jieyi Huang, Chunhao Zhang, Mengyue Wu, and Kenny Zhu, "Transcribing vocal communications of domestic shiba lnu dogs," in *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, Eds., Toronto, Canada, July 2023, pp. 13819–13832, Association for Computational Linguistics.

[3] Theron S. Wang, Xingyuan Li, Chunhao Zhang, Mengyue Wu, and Kenny Q. Zhu, "Phonetic and lexical discovery of canine vocalization," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, Eds., Miami, Florida, USA, Nov. 2024, pp. 13972–13983, Association for Computational Linguistics.

[4] Marius Miron, Sara Keen, Jen-Yu Liu, Benjamin Hoffman, Masato Hagiwara, Olivier Pietquin, Felix Effenberger, and Maddie Cusimano, "Biodenoising: animal vocalization denoising without access to clean data," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[5] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in neural information processing systems*, vol. 33, pp. 3846–3857, 2020.

[6] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin 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.

[7] Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang, "Separate anything you describe," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[8] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] Konstantinos Drossos, Samuel Lipping, and Tuomas 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.

[10] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "Fsd50k," Oct. 2020.

[11] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee 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)*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds., Minneapolis, Minnesota, June 2019, pp. 119–132, Association for Computational Linguistics.

[12] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.

[13] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[14] Karol J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, New York, NY, USA, 2015, MM '15, p. 1015–1018, Association for Computing Machinery.

[15] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John 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.

[16] Simon Rouard, Francisco Massa, and Alexandre 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.

[17] Efthymios Tzinis, Gordon Wichern, Aswin Subramanian, Paris Smaragdis, and Jonathan Le Roux, "Heterogeneous target speech separation," *arXiv preprint arXiv:2204.03594*, 2022.