Course Project Proposal

yp2726 Yueyan Pang
js6626 Jiabao Shen
yt2911 Yuchen Teng

Problem: The project aims to develop a system for language-guided audio source separation, integrating natural language understanding with audio signal processing. The system will use human-provided input to separate audio components from a mixture, focusing on both environmental sounds and music. Applications include audio editing, personalized music experiences, and enhanced speech separation for teleconferencing and assistive listening.

Tools: We will use datasets such as AudioSet and ESC-50, containing paired audio mixtures and natural language descriptions. Python, along with libraries like PyTorch, will be used for model implementation. Deep learning architectures such as ResUNet and Transformer block will be explored, with DSP methods like STFT and Mel-spectrograms for pre-processing. Key challenges include handling language ambiguity and overlapping sounds. Contrastive language-image pre-training model (CLIP), contrastive language-audio pre-training model (CLAP) and multi-modal learning strategies will be used to address these challenges.

Desired Outcomes: We aim to develop a system that can accurately separate target sounds from audio mixtures based on natural language queries, with a focus on both environmental and music sounds. The system should be able to interpret diverse queries and provide reliable separation results, demonstrating its effectiveness in complex auditory environments.

References:

Liu, X., Kong, Q., Zhao, Y., Liu, H., Yuan, Y., Liu, Y., ... & Wang, W. (2023). Separate anything you describe. *arXiv preprint arXiv:2308.05037*.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023, June). 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)* (pp. 1-5). IEEE.

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