Zhisheng Zheng

STATEMENT OF PURPOSE

zszheng@utexas.edu

From a young age, I have always been enthralled by the potential of human communication through audio and speech. This ability to convey complex ideas, emotions, and information intrigued me, but also highlighted numerous unresolved challenges in this domain. These include how to utilize massive unlabeled data to improve recognition accuracy and how to accurately locate sound sources. The advent of artificial intelligence, especially the emergence of large language models (LLMs), has offered a new pathway to address these challenges. This blend of human communication, machine understanding, and the promise of LLMs propelled me into the realm of research. My aim is to bridge the gap between human auditory communication and machine understanding, and to develop Large neural networks applicable to real-world scenarios.

Boosting ASR performance through Self-Supervised and Unsupervised Learning

During my sophomore year, I was astounded by the real-time audio translation feature on Google Recoder, which sparked my curiosity and led me to join the X-LANCE Lab at SJTU under the guidance of Prof. Xie Chen. At that time, self-supervised (SSL) models for Automatic Speech Recognition (ASR) like wav2vec 2.0 and HuBERT, were gaining traction. However, after replicating the results from their respective papers, I found that these models still rely heavily on extensive labeled data for fine-tuning, which seems to contradict the principle of self-supervised learning. In an effort to address this discrepancy, I sought methods to reduce the need for labeled data and discovered wav2vec-U, an unsupervised model that generates coarse pseudo-labels. Leveraging these pseudo-labels, I devised algorithms to filter and isolate high-quality data, significantly reducing the dependency on labeled resources. With this approach, the HuBERT model required only 50 hours of labeled data to match the performance that previously demanded 100 hours. This research was accepted by INTERSPEECH. [1]

In the quest to optimize SSL for ASR, our work, MT4SSL[2], a method that synergizes the strengths of HuBERT and data2vec, not only accelerates model training but also improves convergence. This innovative work was accepted by INTERSPEECH and further honored with a **Best Student Paper nomination**. Moreover, we boosted the HuBERT model's performance by employing pseudo-labeling from wav2vec-U instead of its basic k-means annotations, which INTERSPEECH also

[1] **Zhisheng Zheng**, Z. Ma, Y. Wang, X. Chen, "Unsupervised Active Learning: Optimizing Labeling Cost-Effectiveness for Automatic Speech Recognition," *INTERSPEECH 2023*.

[2] Z. Ma, **Zhisheng Zheng**, C. Tang, Y. Wang, and X. Chen, "MT4SSL: Boosting self-supervised speech representation learning by integrating multiple targets," *INTERSPEECH 2023*.

accepted.[3] Further, to speedup HuBERT's pre-training, we introduced Fast-HuBERT[4], achieving a 5.2x speed increase under similar hardware conditions, a contribution recognized by ASRU.

Sound Source Localization with Large Language Model

My passion for audio and speech led me to collaborate at the Speech, Audio, and Language Technologies (SALT) Lab at UT-Austin with Prof. David Harwath and Prof. Eunsol Choi. To this day, there has been scant exploration in simultaneously addressing the challenges of sound source localization and sound event detection, let alone including distance as a metric. Therefore, we ventured to consider these aspects collectively, integrating the capabilities of large language models (LLM) like Llama 2 to construct a system capable of not only detecting sound events but also discerning the coordinates of the sound source. My current system has shown significant efficacy in identifying sound events within audio and determining their direction and distance with high accuracy (up to 70%). We are preparing to submit our work to the 41st International Conference on Machine Learning (ICML) 2024.

Future Plans

As I embark on my Ph.D. at UT-Austin, I aim to expand my research under the mentorship of Prof. **David Harwath** and Prof. **Eunsol Choi**, delving into the realms of speech processing and large language models. My focus will be on pioneering advancements in localization and navigation, utilizing cutting-edge techniques in large language model. This work promises to not only enhance academic understanding but also to offer practical applications in areas like virtual reality and assistive technologies for individuals with disabilities. In collaboration with these esteemed professors, I am excited to explore the integration of multimodal data to advance the field of embodied AI, aligning with the innovative research culture at UT-Austin.

[3] Z Ma, **Zhisheng Zheng**, G Yang, Y Wang, C Zhang, X Chen, "Pushing the Limits of Unsupervised Unit Discovery for SSL Speech Representation," *INTERSPEECH 2023*.

[4] G Yang, Z Ma, **Zhisheng Zheng**, Y Song, Z Niu, X Chen, "Fast-HuBERT: An Efficient Training Framework for Self-Supervised Speech Representation Learning," *ASRU 2023*.