My research interests lie in natural language processing (NLP), particularly in the realm of spoken language processing. This is motivated by my experience as a second-generation immigrant: I have always struggled to communicate with my family in Taiwan due to my inability to speak Taiwanese Hokkien. Thus, my goal is to enable universal translation. I want to develop technologies that can dismantle language barriers, allowing people to communicate with each other in their own language. **By providing equitable access to technology for all languages, we can reduce communication barriers while maintaining linguistic and cultural diversity.** Universal translation, however, is not a simple task. Due to the lack of training data for all possible language pairs, we need models that can either leverage unlabeled data or perform tasks zero-shot. As such, my research with Professor **Shinji Watanabe** at **Carnegie Mellon University** (CMU) has focused on **developing transparent speech foundation models via large-scale self-supervised or multi-task learning.**

My first project at CMU (and in speech processing) was focused on developing methods to help automatic speech recognition (ASR) models handle the acoustic diversity of multilingual speech across 102 languages. I designed a technique to align the model's representations of each word in the input speech with the identity of the language being spoken. The motivation was that to properly transcribe the speech, the language being spoken has to be first identified, which is easier when the model learns features relevant to each language. The result was a multilingual, multi-task ASR model that achieved state-of-the-art (SOTA) performance on the FLEURS benchmark, **beating the then-leading performance of Google's system by 28.4%**. This paper was published at ICASSP 2023 [1], and won a **Top 3% Paper Award** along with an $850 student travel grant.

My next work targeted self-supervised learning (SSL) with unlabeled data. Leveraging unlabeled speech data is crucial to scaling NLP to more languages, as many do not have a formal writing system. I first did a pilot study on monolingual English SSL and published my first speech foundation model, AR-HuBERT [2], at Interspeech 2023; AR-HuBERT was trained on 60K hours of speech data, and the first SSL speech model from academia that could outperform the SOTA models developed by large technology companies. Most importantly, it was trained only 16 A100 GPUs, while comparable models from industry used 128 GPUs. This work attracted the attention of many researchers, and showed me the importance of building transparent foundation models. People wanted to know how we were able to conduct such large-scale experiments in an academic setting. Sharing that information is crucial in maintaining reproducible research and truly democratizing AI technologies beyond a select few industry labs. **As such, all of my work at CMU in building speech foundation models has open-sourced model weights, training scripts, and training configurations.**

I extended this by developing a multilingual speech foundation model for 137 languages - WavLabLM [3]. **WavLabLM is an SSL speech encoder trained on 40K hours of speech data that achieves comparable performance to Meta's XLS-R that is trained on 436K hours**. I achieved this by focusing on two key aspects: the SSL objective and balancing performance across languages. Multilingual SSL speech encoders have largely focused on end-to-end SSL objectives that use vector quantization to obtain the prediction targets [4]. While this simplifies the training pipeline, we hypothesized that they are less data efficient than models trained with targets obtained offline. The latter method, however, can require significant amounts of tuning to obtain strong offline targets, limiting their use to mostly only English models so far [5]. WavLabLM shows that the heuristics used to obtain the

targets for English models are sufficient for the multilingual setting. I further augmented the objective with a speech de-noising task, making WavLabLM more robust to noisy inputs. Finally, I devised a simple and efficient technique that significantly improves performance on low-resourced languages: just continually pre-train for a few more steps on a language-balanced subset of the data after the first round of pre-training on a large unbalanced dataset. WavLabLM was accepted to ASRU 2023.

I was fortunate to use my experience in training large-scale foundation models in two other works accepted at ASRU: OWSM [6] and YODAS [7]. **OWSM is a transparent reproduction of OpenAI's Whisper with only publicly-accessible data.** I helped scale OWSM to 180K hours of labeled speech data, and we showed that it can achieve similar performance to Whisper across a diverse range of tasks. We upgraded OWSM to support any-to-any speech translation (instead of Whisper's any-to-English), allowing it to perform more zero-shot translation directions. We also alleviate Whisper's hallucination issues by supporting joint CTC/attention decoding. YODAS is the first true large-scale dataset for ASR. We crawled over 500,000 hours of speech across 140 languages from Youtube and aligned it with text transcriptions, leading to over 420,000 hours of labeled data. **This makes YODAS the largest publicly available speech dataset, and is 8 times larger than the next largest labeled corpus.**

All of these speech foundation models are based off of the Transformer architecture, which has limited use in long-context tasks due to the memory inefficiency of self-attention. I investigated this topic during my time as a visiting researcher at **NTT Japan**, where we focused on improving the memory efficiency of speech encoders. **We found that attention could be removed from typical ASR architectures in long-context tasks.** My motivation for proposing this idea was the weighted sum aspect of attention: given a sufficiently long sequence, the weights of each token would be increasingly diluted. If all of the tokens had a similar weight, what was the point of using attention? I verified this idea empirically by comparing to Flash Attention [8] on long-form ASR, where the input sequence during *both training and testing* was up to 10 minutes of audio. The attention-free model achieved almost identical performance to Flash Attention. **I extended this technique to speech SSL, leading to the first speech foundation model that can be used in long-form processing without windowing.**

My current work now focuses on combining all of the lessons and findings from this previous research and using them to scale speech foundation models to previously unseen levels. I am now training SSL models on over 1 million hours of public speech data, which rivals those developed by Google and Meta while remaining accessible and transparent. In the future, I am interested in investigating new methods for SSL pre-training and integrating language modeling techniques for speech generation.

At CMU, I am primarily interested in continuing to work with Shinji Watanabe, who is a renowned speech processing expert that also wishes to extend the technology to more languages. I would also be interested in collaborating with Alex Waibel and Lei Li, who have both worked in speech translation.

References:

[1] William Chen et al. "Improving massively multilingual ASR with auxiliary CTC objectives," *Proc. ICASSP*, 2023.
[2] William Chen et al. "Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute," *Proc. INTERSPEECH*, 2023.
[3] William Chen,et al. "Joint Prediction and Denoising for Large-scale Multilingual Self-supervised Learning," *Proc. ASRU*, 2023.
[4] Arun Babu, et al. "XLS-R: Self-supervised cross-lingual speech representation learning at scale." *arXiv preprint arXiv:2111.09296*, 2021.
[5] Wei-Ning Hsu et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *TASLP*, 2021.
[6] Yifan Peng et al. "Reproducing Whisper-Style Training Using an Open-Source Toolkit and Publicly Available Data," *Proc. ASRU*, 2023.
[7] Xinjian Li et al. "YODAS: Youtube-Oriented Dataset for Audio and Speech," *Proc. ASRU*, 2023.
[8] Tri Dao et al. "FlashAttention: Fast and memory-efficient exact attention with IO-awareness." *Proc NeurIPS*, 2022.