# WILLIAM CHEN

williamchen@cmu.edu ⋄ wanchichen.github.io

## EDUCATION

**Carnegie Mellon University**, Pittsburgh, Pennsylvania *2022-2024*
M.S. in Language Technologies
Language Technologies Institute, School of Computer Science
GPA: 3.94/4.33
Research areas: Speech Processing, Self-Supervised Learning, Foundation Models
Advisor: Shinji Watanabe

**University of Central Florida**, Orlando, Florida *2018-2021*
B.S. with Honors in Computer Science, Magna Cum Laude
B.A. with Honors in History, Cum Laude
Burnett Honors College
GPA: 3.89/4.0

## RESEARCH EXPERIENCE

**Carnegie Mellon University, Audio and Voice Lab** 08.2022 - Present
*Graduate Research Assistant* || Advisor: Dr. Shinji Watanabe

- Working on large-scale AI foundation models for speech processing, leading to 12 co-authored papers published at top speech conferences (pub. [2-6, 8-13, 16]).
- Implemented and trained models for Automatic Speech Recognition (ASR), Speech Translation (ST), and Self-Supervised Learning (SSL). All code and models open-sourced via the ESPnet toolkit.
- Proposed language conditioning technique for multilingual ASR, obtaining state-of-the-art results on the FLEURS benchmark and outperforming prior work from Google by 28.3% (pub. [16]).
- Developed efficient training technique and implementation for speech SSL models, leading to the first large-scale speech SSL model by an academic group (pub. [9]).
- Created WavLabLM, a speech SSL model for 136 languages, achieving comparable performance to Meta's XLS-R, despite training on 10 times less data (pub. [2]).
- Helped created OWSM, a transparent alternative to OpenAI's Whisper for ASR and ST (pub. [6]).

**Llamacha** 01.2022 - Present
*Researcher*

- Helping organize a grassroots initiative towards NLP for indigenous American Languages.
- Co-organized IWSLT 2023, curated a Quechua-Spanish ST dataset for the challenge (pub. [14, 15]).
- Developed QuBERT, a BERT model for Quechua, by creating its largest text corpus (pub. [17]).

**NTT Corporation, Communication Sciences Lab** 05.2023 - 08.2023
*Visiting Researcher* || Advisors: Drs. Marc Delcroix, Takatomo Kano, Atsunori Ogawa

- Worked on speech summarization (SSUM) and long-form speech recognition.
- Developed an open-source toolkit for SSUM that introduces the largest-yet SSUM dataset (pub. [3]).
- Proposed methods to improve memory efficiency of speech encoders, increasing context length from 2 to 30 minutes. Improved performance on ASR and SSUM on the How2 dataset (pub. [20]).
- Invented LongHuBERT, the first modern attention-free speech SSL model, allowing it to be used in long-form speech tasks. State-of-the-art performance on the SLUE-TED SSUM benchmark (pub. [21]).

**University of Central Florida, Computational Biology Lab** 06.2020 - 08.2022
*Undergraduate Research Assistant* || Advisor: Dr. Wei Zhang

- Worked on multi-omics models for cancer sub-type prediction.
- Helped develop a graph neural network that simulates miRNA for gene expression (pub. [1]).

**University of Central Florida, Evolutionary Computation Lab** 01.2020 - 10.2021
*Undergraduate Research Assistant* || Advisor: Dr. Annie Wu

- Worked on using cellular automata to enhance file compression algorithms.

**University of Central Florida, Security and Analytics Lab** 04.2021 - 07.2021
*Undergraduate Research Assistant* || Advisor: Dr. David Mohaisen

- Worked on applying NLP techniques to cybersecurity.
- Curated new dataset by hand-summarizing 1500 security vulnerability reports.
- Fine-tuned T5 on the dataset, showing that it can be used to summarize new reports (pub. [8]).

## WORK EXPERIENCE

**Texas Instruments** 07.2021 - 08.2022
*Software Engineer*

- Full-stack developer on the E-commerce Team that proceed over $1B USD of annual revenue.
- Maintained the company's inventory allocation engine, working in React, Java Spring, and Oracle SQL.
- Upgraded the inventory allocation algorithm to better represent inventory levels.
- Developed performance monitoring framework for inventory engine, reducing support response time.
- Responsible for mentoring one intern and one new-hire.

**uBump.co** 08.2020 - 05.2021
*Chief Information Officer*

- Led front-end development of social media sharing startup. Worked in React and Express.js.
- Helped develop marketing posts on social media, leading to over 2 million views.
- Company was acquired by Bolstered Equity Group for $25K USD.

**Valorantify** 06.2020 - 08.2020
*Software Engineer*

- Front-end developer for one of the first e-sport news and statistics sites for Riot Games' VALORANT.
- Company was acquired by thespike.gg, the second largest VALORANT news site.

**Texas Instruments** 06.2020 - 08.2020
*Software Engineering Intern*

- Developer on Inventory Management team, working in React, Java Spring, and Oracle SQL.
- Created web application to control inventory management engine.

## REFREED PUBLICATIONS, JOURNAL

[1] Khandakar Tanvir Ahmed, Jiao Sun, **William Chen**, Irene Martinez, Sze Cheng, Wencai Zhang, Jeongsik Yong, and Wei Zhang. "In Silico Model for miRNA-mediated Regulatory Network in Cancer". *Briefings in Bioinformatics, Volume 22, Issue 6*, 2021.

## REFREED PUBLICATIONS, CONFERENCE

[2] **William Chen**, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, and Shinji Watanabe. "Joint Prediction and Denoising for Large-scale Multilingual Self-supervised Learning," To appear in *Proc. ASRU*, 2023.

[3] Roshan Sharma, **William Chen**, Takatomo Kano, Ruchira Sharma, Atsunori Ogawa, Siddhant Arora, Marc Delcroix, Rita Singh, Shinji Watanabe, Bhiksha Raj. "ESPNet-SUMM: Introducing a novel large dataset, toolkit, and a cross-corpora evaluation of speech summarization systems," To appear in *Proc. ASRU*, 2023.

[4] Jiatong Shi, **William Chen**, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang et al. "Findings of the 2023 ML-SUPERB Challenge: Pre-Training and Evaluation over More Languages and Beyond," To appear in *Proc. ASRU*, 2023.

[5] Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, **William Chen**, Sayaka Shiota, Shinji Watanabe. "YODAS: Youtube-Oriented Dataset for Audio and Speech," To appear in *Proc. ASRU*, 2023.

[6] Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, **William Chen**, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-weon Jung, Soumi Maiti, Shinji Watanabe. "Reproducing Whisper-Style Training Using an Open-Source Toolkit and Publicly Available Data," To appear in *Proc. ASRU*, 2023.

[7] Takatomo Kano, Atsunori Ogawa, Marc Delcroix, Kohei Matsuura, Takanori Ashihara, **William Chen**, Shinji Watanabe. "Summarize while Translating: Universal Model with Parallel Decoding for Summarization and Translation," To appear in *Proc. ASRU*, 2023.

[8] Hattan Althebeiti, Brett Fazio, **William Chen**, David Mohaisen. "Mujaz: A Summarization-based Approach for Normalized Vulnerability Description," *Proc. ACM CCS*, 2023.

[9] **William Chen**, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, and Shinji Watanabe. "Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute," *Proc. INTERSPEECH*, 2023.

[10] Jiyang Tang, **William Chen**, Xuankai Chang, Shinji Watanabe, Brian MacWhinney. "A New Benchmark of Aphasia Speech Recognition and Detection Based on E-Branchformer and Multi-task Learning," *Proc. INTERSPEECH*, 2023.

[11] Jiatong Shi, Dan Berrebbi, **William Chen**, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang et al. "ML-SUPERB: Multilingual Speech Universal PERformance Benchmark," *Proc. INTERSPEECH*, 2023.

[12] Yifan Peng, Kwangyoun Kim, Felix Wu, Brian Yan, Siddhant Arora, **William Chen**, Jiyang Tang, Suwon Shon, Prashant Sridhar, and Shinji Watanabe. "A Comparative Study on E-Branchformer vs Conformer in Speech Recognition, Translation, and Understanding Tasks," *Proc. INTERSPEECH*, 2023.

[13] Brian Yan, Jiatong Shi, Soumi Maiti, **William Chen**, Xinjian Li, Yifan Peng, Siddhant Arora, Shinji Watanabe. "CMU's IWSLT 2023 Simultaneous Speech Translation System," *Proc. IWSLT*, 2023.

[14] John E. Ortega, Rodolfo Zevallos, **William Chen**. "QUESPA Submission for the IWSLT 2023 Dialect and Low-resource Speech Translation Tasks," *Proc. IWSLT*, 2023.

[15] Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, **William Chen**, Khalid Choukri, et al.. "Findings of the IWSLT 2023 Evaluation Campaign," *Proc. IWSLT*, 2023.

[16] **William Chen**, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. "Improving massively multilingual asr with auxiliary ctc objectives," *Proc. ICASSP*, 2023.

[17] Rodolfo Zevallos, John Ortega, **William Chen**, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. "Introducing QuBERT: A Large Monolingual Corpus and BERT Model for Southern Quechua," *Proc. DeepLo*, 2022.

[18] **William Chen** and Brett Fazio. "Morphologically-guided Segmentation for Translation of Low-Resource Agglutinative Languages," *Proc. LoResMT*, 2021.

[19] **William Chen** and Brett Fazio. "The UCF Systems for the LoResMT 2021 Machine Translation Shared Task," *Proc. LoResMT*, 2021.

## UNPUBLISHED MANUSCRIPTS

[20] **William Chen**, Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. "Train Long and Test Long: Leveraging Full Document Contexts in Speech Processing."

[21] **William Chen**, Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. "LongHuBERT: Evaluating the Importance of Attention in Self-supervised Speech Encoders."

[22] Jee-weon Jung, Roshan Sharma, **William Chen**, Bhiksha Raj, and Shinji Watanabe. "AugSumm: Towards Generalizable Speech Summarization Using Synthetic Labels from Large Language Models."

[23] Siddhant Arora, Roshan Sharma, Ankita Pasad, Hira Dhamyal, **William Chen**, Suwon Shon, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. "SLUE-PERB: A Spoken Language Understanding Performance Benchmark and Toolkit."

## FUNDING, AWARDS AND HONORS

| | |
|---|---|
| **Monte Jade SE Innovation Competition (2023)** | $5000 1st place entrepreneurship award |
| **IEEE SPS Student Travel Grant (2023)** | $850 award for ICASSP 2023 [16] |
| **ICASSP Top 3% Paper Award (2023)** | Top paper award at ICASSP 2023 [16] |
| **CMU LTI Research Fellowship (2023)** | Full funding for master's degree at CMU |
| **FLORES 101 Compute Grant (2021)** | $750 award in Azure credits |
| **LoResMT Best Paper Honorable Mention (2021)** | Top paper award at LoResMT 2021 [18] |
| **NSF REU Scholarship (2020)** | Funded undergraduate research at UCF |
| **Benaquisto Scholarship (2018)** | Fully-funded merit scholarship |
| **Bright Futures Scholarship (2018)** | Full-tuition merit scholarship |
| **National Merit Finalist (2018)** | Awarded to top 1% of PSAT scorers |

## SERVICE

### Reviewer

· LREC (2024), ICASSP (2024), IWSLT (2023), CoCo4MT (2022, 2023), ALTNLP (2022), NTTT (2022)

### Co-Organizer

· IWSLT (2023, 2024), CoCo4MT (2022, 2023), ALTNLP (2022)

### Volunteer

· ACL-IJCNLP (2021)

## SKILLS

| | |
|---|---|
| **Languages** | English (native), Mandarin Chinese (native), French (fluent) |
| **Programming Languages** | Python, Java, Javascript, Typescript, C, C#, Rust |
| **Frameworks** | PyTorch, Tensorflow, Next.js, React, Express, Actix-Web |