WILLIAM CHEN

williamchen@cmu.edu ⋄ wanchichen.github.io

EDUCATION

Carnegie Mellon University, Pittsburgh, Pennsylvania
PhD in Language Technologies 2024 - Present
M.S. in Language Technologies 2022 - 2024
Language Technologies Institute, School of Computer Science
GPA: 4.05
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

RESEARCH EXPERIENCE

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

· Large-scale AI foundation models for speech processing.

Llamacha 01.2022 - Present
Researcher

· A grassroots initiative towards NLP for indigenous American Languages.

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

· Speech summarization and long-form speech recognition.

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

· Multi-omics models for cancer sub-type prediction.

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

· Applying NLP techniques to cybersecurity.

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.

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.

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**, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, Shinji Watanabe. "Towards Robust Speech Representation Learning for Thousands of Languages" *Proc. EMNLP*, 2024. **Best Paper Award**.

[3] **William Chen**, Brian Yan, Chih-Chen Chen, Shinji Watanabe. "FLORAS 50: A Massively Multilingual Multitask Benchmark for Long-form Conversational Speech" *Proc. SLT*, 2024.

[4] Masao Someki, Kwanghee Choi, Siddhant Arora, **William Chen**, Samuele Cornell, Jionghao Han, Yifan Peng, Jiatong Shi, Vaibhav Srivastav, Shinji Watanabe. "ESPnet-EZ: Python-only ESPnet for Easy Fine-tuning and Integration" *Proc. SLT*, 2024.

[5] Jiatong Shi, Jinchuan Tian, Yihan Wu, Jee-weon Jung, Jia Qi Yip, Yoshiki Masuyama, **William Chen** et al.. "ESPnet-Codec: Comprehensive Training and Evaluation of Neural Codecs for Audio, Music, and Speech" *Proc. SLT*, 2024.

[6] Xi Xu, Siqi Ouyang, Brian Yan, Patrick Fernandes, **William Chen**, Lei Li, Graham Neubig, Shinji Watanabe. "CMU's IWSLT 2024 Simultaneous Speech Translation System" *Proc. IWSLT*, 2024.

[7] Brian Yan, Patrick Fernandes, Jinchuan Tian, Siqi Ouyang, **William Chen**, Karen Livescu, Lei Li, Graham Neubig, Shinji Watanabe. "CMU's IWSLT 2024 Offline Speech Translation System: A Cascaded Approach For Long-Form Robustness" *Proc. IWSLT*, 2024.

[8] Ibrahim Sa'id Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, **William Chen**, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kim Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John Philip McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, Rodolfo Zevallos "Findings of the IWSLT 2024 Evaluation Campaign" *Proc. IWSLT*, 2024.

[9] Tejes Srivastava, Jiatong Shi, **William Chen**, Shinji Watanabe. "EFFUSE: Efficient self-supervised feature fusion for E2E ASR in multilingual and low resource scenarios." *Proc. INTERSPEECH*, 2024. **Best Paper Award**.

[10] Yifan Peng, Jinchuan Tian, **William Chen**, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee-weon Jung, Shinji Watanabe. "OWSM v3.1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer." *Proc. INTERSPEECH*, 2024.

[11] Jiatong Shi, Shih-Heng Wang*, **William Chen***, Martijn Bartelds*, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung-yi Lee, Shinji Watanabe. "ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets." *Proc. INTERSPEECH*, 2024.

[12] Jinchuan Tian, Yifan Peng, **William Chen**, Kwanghee Choi, Karen Livescu, Shinji Watanabe. "On the Effects of Heterogeneous Data Sources on Speech-to-Text Foundation Models." *Proc. INTERSPEECH*, 2024.

[13] Siddhant Arora, Ankita Pasad, Chung-Ming Chien, Jionghao Han, Roshan Sharma, Jee-weon Jung, Hira Dhamyal, **William Chen**, Suwon Shon, Hung-yi Lee, Karen Livescu, Shinji Watanabe. "On the Evaluation of Speech Foundation Models for Spoken Language Understanding." *Findings of ACL*, 2024.

[14] Chih-Chen Chen*, **William Chen***, Rodolfo Zevallos, John E. Ortega. "Evaluating Self-Supervised Speech Representations for Indigenous American Languages." *Proc. LREC-COLING*, 2024.

[15] **William Chen**, Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. "Train Long and Test Long: Leveraging Full Document Contexts in Speech Processing." *Proc. ICASSP*, 2024.

[16] Jee-weon Jung, Roshan Sharma, **William Chen**, Bhiksha Raj, and Shinji Watanabe. "AugSumm: Towards Generalizable Speech Summarization Using Synthetic Labels from Large Language Models." *Proc. ICASSP*, 2024.

[17] **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," *Proc. ASRU*, 2023.

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

[19] Jiatong Shi, **William Chen**, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, Shinji Watanabe. "Findings of the 2023 ML-SUPERB Challenge: Pre-Training and Evaluation over More Languages and Beyond," *Proc. ASRU*, 2023.

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

[21] 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," *Proc. ASRU*, 2023.

[22] 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," *Proc. ASRU*, 2023.

[23] Ananya Ganesh, Marine Carpuat, **William Chen**, Katharina Kann, Constantine Lignos, John E Ortega, Jonne Sälevä, Shabnam Tafreshi, Rodolfo Zevallos. "Findings of the CoCo4MT 2023 Shared Task on Corpus Construction for Machine Translation." *Proc. AMTA*, 2023.

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

[25] **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.

[26] 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.

[27] Jiatong Shi, Dan Berrebbi, **William Chen**, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, Shinji Watanabe. "ML-SUPERB: Multilingual Speech Universal PERformance Benchmark," *Proc. INTERSPEECH*, 2023.

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

[29] 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.

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

[31] Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, **William Chen**, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, Rodolfo Zevallos. "Findings of the IWSLT 2023 Evaluation Campaign," *Proc. IWSLT*, 2023.

[32] **William Chen**, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, Shinji Watanabe. "Improving massively multilingual asr with auxiliary ctc objectives," *Proc. ICASSP*, 2023. **Top 3% Paper Award, IEEE SPS Student Travel Grant Award**

[33] John E Ortega, Marine Carpuat, **William Chen**, Katharina Kann, Constantine Lignos, Maja Popović, Shabnam Tafreshi. "Proceedings of the Workshop on Corpus Generation and Corpus Augmentation for Machine Translation". *Proc. AMTA*, 2022.

[34] 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.

[35] **William Chen** and Brett Fazio. "Morphologically-guided Segmentation for Translation of Low-Resource Agglutinative Languages," *Proc. LoResMT*, 2021. **Best Paper Award Honorable Mention.**

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

UNPUBLISHED MANUSCRIPTS

[37] Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, **William Chen**, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Chih-Kai Yang, Fabian Ritter-Gutierrez, Ming To Chuang, Kuan-Po Huang, Siddhant Arora, You-Kuan Lin, Eunjung Yeo, Kalvin Chang, Chung-Ming Chien, Kwanghee Choi, Cheng-Hsiu Hsieh, Yi-Cheng Lin, Chee-En Yu, I Chiu, Heitor R Guimarães, Jionghao Han, Tzu-Quan Lin, Tzu-Yuan Lin, Homu Chang, Ting-Wu Chang, Chun Wei Chen, Shou-Jen Chen, Yu-Hua Chen, Hsi-Chun Cheng, Kunal Dhawan, Jia-Lin Fang, Shi-Xin Fang, Kuan-Yu Fang Chiang, Chi An Fu, Hsien-Fu Hsiao, Ching Yu Hsu, Shao-Syuan Huang, Lee Chen Wei, Hsi-Che Lin, Hsuan-Hao Lin, Hsuan-Ting Lin, Jian-Ren Lin, Ting-Chun Liu, Li-Chun Lu, Tsung-Min Pai, Ankita Pasad, Shih-Yun Shan Kuan, Suwon Shon, Yuxun Tang, Yun-Shao Tsai, Jui-Chiang Wei, Tzu-Chieh Wei, Chengxi Wu, Dien-Ruei Wu, Chao-Han Huck Yang, Chieh-Chi Yang, Jia Qi Yip, Shao-Xiang Yuan, Vahid Noroozi, Zhehuai Chen, Haibin Wu, Karen Livescu, David Harwath, Shinji Watanabe, Hung-yi Lee. "Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks"

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

[39] John E. Ortega , **William Chen**, Ibrahim Said Ahmad. "Nollywood: Let's Go to the Movies!"

FUNDING, AWARDS AND HONORS

EMNLP Best Paper Award (2024)	Overall Best Paper Award [2]
Interspeech Best Paper Award (2024)	Overall Best Paper Award [9]
Monte Jade SE Innovation Competition (2024)	$3000 1st place entrepreneurship award
Monte Jade SE Innovation Competition (2023)	$5000 1st place entrepreneurship award
IEEE SPS Student Travel Grant (2023)	$850 award for ICASSP 2023 [32]
ICASSP Top 3% Paper Award (2023)	Top paper award at ICASSP 2023 [32]
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 [36]
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

- ICASSP (2024, 2025), ACL (2024), EMNLP (2024)
- LREC-COLING (2024), COLING (2025), IWSLT (2023, 2024)
- CoCo4MT (2022, 2023), ALTNLP (2022), NTTT (2022)

Organizer

· ASRU 2023 ML-SUPERB Challenge, Interspeech 2025 ML-SUPERB Challenge
· IWSLT (2023, 2024), CoCo4MT (2022, 2023), ALTNLP (2022)

Volunteer

· ACL-IJCNLP (2021)

TEACHING

18-781/11-751 Speech Recognition and Understanding 08.2024 - 12.2024
Graduate Teaching Assistant || Instructor: Prof. Shinji Watanabe

SKILLS

Languages	English (native), Mandarin Chinese (native)
Programming	Python, Java, Javascript, Typescript, React
Machine Learning	PyTorch, ESPnet, fairseq, HuggingFace