Yuan Gong

Member of Technical Staff, xAI yuangongfdu@gmail.com · https://yuangongnd.github.io

Research Interest

AI for Audio, Speech, and Natural Language Processing; Large Language Models; AI for Health; Secure AI

Employment

xAI Corp., Palo Alto CA — Aug 2024 - Present
The Company's First Hire in the Audio/Speech Domain, Manager: Elon Musk
Member of Technical Staff

Massachusetts Institute of Technology, Cambridge MA
CSAIL Spoken Language Systems Group, Host: Dr. James Glass
Research Scientist II — Aug 2023 - Aug 2024
Postdoctoral Associate — Aug 2020 - July 2023

Amazon Web Service, Seattle WA
Comprehensive Medical Team
Applied Research Scientist Intern — May 2019 - Aug 2019

University of Notre Dame, Notre Dame IN
Mobile Computing Lab, Advisor: Prof. Christian Poellabauer
Research Assistant — Aug 2015 - Jul 2020

Philips Healthcare, Shanghai, China
Summer Intern — Jul 2012 - Aug 2012

Education

University of Notre Dame, Notre Dame IN, USA
Ph.D. and M.S. in Computer Science and Engineering — 2015-2020
Advisor: Prof. Christian Poellabauer, GPA: 4.0/4.0
Thesis: Healthcare Applications and Security Concerns of Speech Processing Systems.

Fudan University, Shanghai, China
B.S. in Electronic Engineering (Biomedical Engineering Major) — 2011-2015

Awards

ASRU 2023 Best Paper Finalist (top 3% paper, 12/435) — December 2023
ICASSP 2023 Outstanding Reviewer — June 2023
Interspeech 2019 Best Student Paper Award Nomination — Jul 2019
AVEC 2017 Depression Detection Challenge Winner — Oct 2017
Fudan First Prize Scholarship (top 3%) and Outstanding Graduates — Apr/Jul 2015

Five Representative Papers (My 30 publications have over 3,300 citations, according to Google Scholar)

1. Yuan Gong, Hongyin Luo, Alex Liu, Leonid Karlinsky, James Glass, **Listen, Think, and Understand**, ICLR 2024. (the first audio large language model, invited talk at MIT EI Seminar and 2023 SANE workshop)

2. Yuan Gong, Andrew Rouditchenko, Alex Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, James Glass, **Contrastive Audio-Visual Masked Autoencoder**, ICLR 2023. (top-25% paper, covered by MIT News)

3. Yuan Gong, Yu-An Chung, James Glass, **AST: Audio Spectrogram Transformer**, Interspeech 2021. (1,100+ citations, 3rd most cited Interspeech 2021 paper, 1 million+ downloads/month from Hugging Face)

4. Yuan Gong, Jian Yang, Jacob Huber, Mitchell MacKnight, Christian Poellabauer, **ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems**, Interspeech 2019. (best student paper nomination)

5. Yuan Gong and Christian Poellabauer, **Topic Modeling Based Multi-modal Depression Detection**, The 7th Audio/Visual Emotion Challenge and Workshop (AVEC) in conjunction with ACM Multimedia (ACM-MM), 2017. (depression detection challenge winner)

Professional Service

Co-organizer

IEEE Spoken Language Technology Workshop 2024 Challenge: Generative Speech Transcription Error Correction in Diverse Scenarios Challenge 2024

Interspeech 2024 Special Session: Responsible Speech Foundation Models 2024

Reviewer Sep 2018 - present

Reviewed over 70 papers for IEEE TASL, TPAMI, TDSC, THMS, SPL, Springer Machine Learning, Interspeech, DCASE, ICASSP, ICCV, NeurIPS, etc.

ACM-BCB 2019 Travel Grant Committee Member Aug 2019

Dissertation Committee Member of Marisa Cameron (Master Student) Apr 2017

Invited Talks and Guest Lectures

From Audio Perception to Understanding: A Path Towards Audio AGI (1.5-hour talk)

Amazon GenAI 4/24/2024

General Audio Processing: Perception, Understanding, and Generation (Guest Lecture)

MIT MIT 6.8620/HST.728 Spoken Language Processing 4/18/2024

How We Evaluate Our Audio and Speech Large Language Model?

ASRU Workshop on Speech Foundation Models and their Performance Benchmarks 12/16/2023

Recent Progress of MIT SLS's Research on Audio Classification and Understanding (two-hour tutorial talk)

Amazon Alexa 12/15/2023

Audio Large Language Models: From Sound Perception to Understanding

Speech and Audio in the Northeast Workshop (SANE 2023) 10/26/2023

MIT Embodied Intelligence Seminar 10/19/2023

Contrastive Audio-Visual Masked Autoencoder

Hong Kong University of Science and Technology (Guangzhou) 10/11/2023

IBM Watson AI Lab 7/28/2023

Large Language Models that Listen

Signify Research 7/21/2023

Takeda 5/30/2023

Introduction of Audio Spectrogram Transformer - Architecture, Training, and Pre-training

Adobe Research 7/12/2022

ByteDance 6/14/2022

Mitsubishi Electric Research Laboratories 6/8/2022

AI Time 5/26/2022

MIT Embodied Intelligence Seminar 10/14/2021

ISCA SIGML Seminar 6/16/2021

General Audio Processing (Guest Lecture)

MIT 6.345/HST.728 Spoken Language Processing 4/19/2022

Win the Cat and Mouse Game: Ensuring the Security of the Speech Processing Systems to Real World Threats (Guest Lecture)

Notre Dame CSE60641 Graduate Operating Systems 10/31/2019

Speech Processing: Machine Learning Approaches, Novel Applications, and New Security Concerns (Guest Lecture)

Notre Dame CSE60641 Graduate Operating Systems 9/20/2018

Conference Papers (conference ranks based on h5-index, according to Google Scholar Metrics)

1. Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, James Glass, **Listen, Think, and Understand**, Proceedings of the 12th International Conference on Learning Representations (ICLR 2024, **the first audio large language model, invited talk at MIT EI Seminar and 2023 SANE workshop**).
[ICLR ranks #2 in Artificial Intelligence]
2. Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinksy, Hilde Kuehne, Rogerio Feris, James Glass, **Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation**, Proceedings of the 25th Conference of the International Speech Communication Association (Interspeech 2024).
[Interspeech ranks #2 in Acoustic and Sound, and #4 in Signal Processing]
3. Liming Wang, Yuan Gong, Nauman Dawalatabad, Marco Vilela, Katerina Placek, Brian Tracey, Yishu Gong, Alan Premasiri, Fernando Vieira, James Glass, **Automatic Prediction of Amyotrophic Lateral Sclerosis Progression using Longitudinal Speech Transformer**, Proceedings of the 25th Conference of the International Speech Communication Association (Interspeech 2024).
[Interspeech ranks #2 in Acoustic and Sound, and #4 in Signal Processing]
4. Tianhua Zhang, Jiaxin Ge, Hongyin Luo, Yung-Sung Chuang, Mingye Gao, Yuan Gong, Xixin Wu, Yoon Kim, Helen Meng, and James Glass, **Natural Language Embedded Programs for Hybrid Language Symbolic Reasoning**, Proceedings of Findings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Findings of NAACL 2024).
[NAACL ranks #3 in Computational Linguistics]
5. Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Helen Meng, and James Glass, **Search Challenges Large Language Models**, Proceedings of Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP 2023).
[EMNLP ranks #2 in Computational Linguistics]
6. Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass, **Joint Audio and Speech Understanding**, 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2023, **top 3% paper (12/435), best paper finalist**).
7. Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass, **Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers**, Proceedings of the 24th Conference of the International Speech Communication Association (Interspeech 2023).
[Interspeech ranks #2 in Acoustic and Sound, and #4 in Signal Processing]
8. Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass, **Contrastive Audio-Visual Masked Autoencoder**, Proceedings of the 11th International Conference on Learning Representations (ICLR 2023, **notable-top-25% paper, covered by MIT News**).
[ICLR ranks #2 in Artificial Intelligence]
9. Jian Yang, Bryan Xia, John Bailey, Yuan Gong, John Templeton, and Christian Poellabauer, **Improving Computational Efficiency of Voice Anti-Spoofing Models**, Proceedings of the 20th IEEE International Conference on Mobile Ad-Hoc and Smart Systems (MASS 2023).
10. Nauman Dawalatabad, Yuan Gong, Sameer Khurana, Rhoda Au, and James Glass, **Detecting Dementia from Long Neuropsychological Interviews**, Proceedings of Findings of the 2022 Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP 2022).
[EMNLP ranks #2 in Computational Linguistics]
11. Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass, **Transformer-Based Multi-Aspect Multi-Granularity Non-native English Speaker Pronunciation Assessment**, International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2022).
[ICASSP ranks #1 in Acoustic and Sound, and #3 in Signal Processing]
12. Yuan Gong, Jin Yu, and James Glass, **VocalSound: A Dataset for Improving Human Vocal Sounds Recognition**, International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2022).

[ICASSP ranks #1 in Acoustic and Sound, and #3 in Signal Processing]

13. Yuan Gong, Cheng-I Jeff Lai, Yu-An Chung, and James Glass, **SSAST: Self-Supervised Audio Spectrogram Transformer**, The 36th AAAI Conference on Artificial Intelligence (AAAI 2022).
[AAAI ranks #4 in Artificial Intelligence]
14. Yuan Gong, Yu-An Chung, and James Glass, **AST: Audio Spectrogram Transformer**, Proceedings of the 22nd Conference of the International Speech Communication Association, (Interspeech 2021, **3rd most cited paper among 963 Interspeech 2021 papers, 35k+ model downloads/month**).
[Interspeech ranks #2 in Acoustic and Sound, and #4 in Signal Processing]
15. Yuan Gong, Boyang Li, Christian Poellabauer, Yiyu Shi, **Real-time Adversarial Attacks**, The 28th International Joint Conference on Artificial Intelligence (IJCAI 2019).
[IJCAI ranks #9 in Artificial Intelligence]
16. Yuan Gong, Jian Yang, Jacob Huber, Mitchell MacKnight, and Christian Poellabauer, **ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems**, Proceedings of the 20th Conference of the International Speech Communication Association (Interspeech 2019, **best student paper nomination**).
[Interspeech ranks #2 in Acoustic and Sound, and #4 in Signal Processing]
17. Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer, **Non-local Second-order Attention Networks for Person Re-identification**, International Conference on Computer Vision (ICCV 2019).
[ICCV ranks #3 in Computer Vision and Pattern Recognition]
18. Yuan Gong and Christian Poellabauer, **Impact of Aliasing on Deep CNN-Based End-to-End Acoustic Models**, Proceedings of the 19th Conference of the International Speech Communication Association (Interspeech 2018).
[Interspeech ranks #2 in Acoustic and Sound, and #4 in Signal Processing]
19. Yuan Gong, Hasini Yatawatte, Christian Poellabauer, Sandra Schneider, and Susan Latham, **Automatic Autism Spectrum Disorder Detection Using Everyday Vocalizations Captured by Smart Devices**, The 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2018).
20. Yuan Gong and Christian Poellabauer, **Protecting Voice Controlled Systems Using Sound Source Identification Based on Acoustic Cues**, The 27th International Conference on Computer Communications and Networks (ICCCN 2018).
21. Yuan Gong and Christian Poellabauer, **Continuous Assessment of Children's Emotional States using Acoustic Analysis**, The 5th IEEE International Conference on Healthcare Informatics (ICHI), 2017.

Journal Papers (journal ranks based on h5-index, according to Google Scholar Metrics)

1. Yuan Gong, Sameer Khurana, Andrew Rouditchenko, and James Glass, **CMKD: CNN/Transformer-Based Cross-Model Knowledge Distillation for Audio Classification**, in submission to IEEE Transactions on Pattern Analysis and Machine Intelligence (first round decision: major revision).
2. Yuan Gong, Alexander H. Liu, Andrew Rouditchenko, and James Glass, **UAVM: Towards Unifying Audio and Visual Models**, IEEE Signal Processing Letters, 2022. [Impact Factor=3.9]
3. Yuan Gong, Yu-An Chung, James Glass, **PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation**, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.
[TASLP ranks #1 in Acoustic and Sound Journals, Impact Factor=5.4]
4. Yuan Gong, Jian Yang, Christian Poellabauer, **Detecting Replay Attacks Using Multi-Channel Audio: A Neural Network-Based Method**, IEEE Signal Processing Letters, 2020. [Impact Factor=3.9]
5. Yuan Gong, Jin Cao, Zehui Luo, and Guohui Zhou, **A Smart Low-Power-Consumption ECG Monitor Based on MSP430F5529 and CC2540**, Chinese Journal of Medical Instrumentation, 2015 (**TI national (China) biomedical device design contest winner**).

Workshop Papers and Posters

1. Yuan Gong and Christian Poellabauer, **Crafting Adversarial Examples For Speech Paralinguistics Applications**, The DYNAMICS Workshop in conjunction with the Annual Computer Security Applications Conference (ACSAC 2018).
2. Yuan Gong, Kevin Shin, and Christian Poellabauer, **Improving LIWC Using Soft Word Matching (Poster)**, The 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2018).
3. Yuan Gong and Christian Poellabauer, **An Overview of Vulnerabilities of Voice Controlled Systems**, 1st International Workshop on Security and Privacy for the Internet-of-Things, 2018.
4. Yuan Gong and Christian Poellabauer, **Topic Modeling Based Multi-modal Depression Detection**, The 7th Audio/Visual Emotion Challenge and Workshop (AVEC) in conjunction with ACM Multimedia (ACM-MM), 2017. **(depression detection challenge winner)**

Open-Source Software (Over 2,400 Stars at GitHub)

1. Audio Spectrogram Transformer (1005 stars, 187 forks)
2. Self-Supervised Audio Spectrogram Transformer (340 stars, 54 forks)
3. Listen, Think, and Understand (286 stars, 18 forks)
4. Whisper-AT (259 stars, 22 forks)
5. Contrastive Audio-Visual Masked Autoencode (201 stars, 20 forks)
6. PSLA Audio Classification Training Pipeline (124 stars, 15 forks)
7. Transformer-Based Pronunciation Assessment. (120 stars, 24 forks)
8. VocalSound Dataset (91 stars, 10 forks)
9. Unified Audio and Visual Model (53 stars, 2 forks)
10. ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems (32 stars, 2 forks)
11. Real-Time Adversarial Attacks (19 stars, 3 forks)
12. Neural Network Based Multi-channel Audio Antispoofing

Media Coverage

Scaling Audio-Visual Learning without Labels, MIT News, June 2023
https://news.mit.edu/2023/scaling-audio-visual-learning-without-labels-0605

AK Daily Paper Tweets (244.5K Followers, now Hugging Face Daily Papers)
AST: Audio Spectrogram Transformer (April 2021)
https://twitter.com/_akhaliq/status/1379237749471993856
SSAST: Self-Supervised Audio Spectrogram Transformer (October 2021)
https://twitter.com/_akhaliq/status/1450634611625693184
Vocalsound: A Dataset for Improving Human Vocal Sounds Recognition (May 2022)
https://twitter.com/_akhaliq/status/1523857971691888642
Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers (June 2023)
https://twitter.com/_akhaliq/status/1677150590516834305

Students

PhD Students Collaborated with as Postdoc/Research Scientist:
Yu-An Chung, Cheng-I Lai, Sameer Khurana, Alexander H. Liu, Andrew Rouditchenko (all at MIT).

Mentored Master's Students for Research:
Marisa Cameron (2016-2017, on the master's committee), Yu Jiang (2020-2021).

Undergraduate Students Mentored for Research for at Least One Semester:
Michael Parowski (2015 Fall), Jorge Diaz-Ortiz (2016 Fall), John Considine (2017 Spring), Kevin Shin (2017 Fall, 2018 Spring), Royce Branning (2018 Spring), Jacob Huber (2018 Fall), Mitchell MacKnight (2018 Fall), Jorge Jose Daboub Silhy (2019 Fall), John Bailey (2019 Fall) (all at Notre Dame).