Research Statement of Yuan Gong

1 Objective

I am an audio, speech, and natural language processing researcher. Automatic audio and speech processing has evolved rapidly since the advent of *Audrey*, the first automatic speech recognition system 70 years ago. Indeed, auditory ability is foundational, and we expect machines to eventually exhibit capabilities similar to our own. Enabling machines to *listen* and *speak* not only benefits numerous applications by allowing machines to interact with the physical world, but it also helps break down communication barriers among people. Throughout my previous research endeavors, my expertise has covered a broad spectrum from algorithm development to applications. Looking forward, I am eager to dedicate myself to the development of **Audio and Speech Artificial General Intelligence (Audio and Speech AGI) that is not only smarter with a more generalized scope of understanding and problem-solving capabilities, but also safer and more socially beneficial**, particularly in the following aspects:

First, most existing audio and speech AI systems are *specialized*, meaning they're designed for limited tasks with limited *generalization* capability. In pursuit of artificial general intelligence, my goal is to develop a **universal audio and speech foundation model** capable of perceiving, understanding, and generating sounds. This model would handle a broad, and eventually complete, array of audio/speech tasks and seamlessly incorporate with other modalities such as vision, paralleling human intelligence.

Second, with the rapid advance of audio and speech AI, the potential *negative* social impacts it brings cannot be neglected. For instance, modern text-to-speech systems can clone a voice with just a 3-second sample, and generative AI might produce work that infringes on copyright. Such capabilities pose serious risks when misused. Thus, I aim to develope audio and speech AGI that is **safe, trustworthy, and copyright-compliant**, incorporating technologies like fake detection, anonymization, and watermarking.

Third, Audio and Speech AGI could also bring positive impacts on healthcare. Speech carries vast information, including indicators of the speaker's health, and can serve as a non-invasive, convenient, and low-cost biomarker. My aim is to create **speech-based systems for early disease detection**, particularly for neuropsychological disorders such as dementia and Parkinson's, and bring them from lab to real-world to benefit millions of people who have limited access to medical resources.

I believe Audio and Speech AGI represents the future of audio and speech processing, and would align with the department's long-term goals and the interests of its faculty. I also recognize developing AGI is ambitious and could be challenging. Therefore, I am keen to engage in close collaboration with other faculty members to make the department a pioneer in Audio and Speech AGI. In the past, I assisted my advisors in writing seven grant proposals, four of which secured funding, including those from industry and federal agencies. In the future, I intend to proactively seek funding from diverse sources, e.g., NSF-IIS for audio understanding and generation research, NSF-SaTC for secure and trustworthy audio AI research, and NIH for health-related speech AI research. The alignment of AGI with major tech companies' interests also presents opportunities like Google Research Scholar and Amazon Research Awards. I also plan to work closely with other faculty members and external partners on joint projects.

2 Past Research

2.1 Foundational Audio AI Algorithms

Learning algorithms are the foundational pillar in the evolution of automatic audio and speech AI, which has matured from the early stages of template matching, through the adoption of hidden Markov models, to convolutional neural networks (CNNs). At the forefront of this progression, I have introduced the Audio Spectrogram Transformers (AST) [1], the first purely attention-based approach. AST allows the unification of audio, visual, and language modalities with a *single* Transformer model. Building on this work, I further developed a self-supervised AST [2] and a new training framework [3], alongside new algorithms for joint audio event and speech recognition [4].

AST has replaced CNNs as the de facto standard model for audio classification. It is included in the Hugging Face Transformers, a leading deep learning library, and is one of only two audio models developed in academia to be included. Even after 2.5 years since its release, AST maintains a high download rate with about **35,000 downloads** each month (as of November 2023)[1]. The AST paper has been **cited over 500 times**, and is the third-most cited paper out of all 963 papers presented at Interspeech 2021, a top-tier conference in the field of audio and speech research.

2.2 Audio and Speech Large Language Models

Existing machine learning systems, while being strong in auditory perception, possess very limited *reasoning* and *understanding* capabilities. For example, the model may recognize a clock chime 6 times, but not know that it indicates a time of 6 o'clock. On the other hand, pure language modern large language models (LLMs) cannot listen but possess strong audio knowledge and reasoning abilities. The potential synergy between conventional audio models and LLMs in sound perception and reasoning motivates me to integrate them into a single model (called LTU) that can *listen* to, *think* about, and *understand* the sound environment [5]. On top of LTU, I further developed a model called LTU-AS [6] that can jointly understand audio events, spoken text, and paralinguistics.

LTU and LTU-AS, as **the first generation audio large language models**, garnered immediate attention upon release. Following this, I was honored with invitations to deliver five talks, including hour-long talks at the MIT Embodied AI Seminar[2] and the Speech and Audio in the Northeast Workshop (SANE 2023)[3], where the other speakers are mainly established professors.

2.3 Connect Audio and Vision Modalities

Acoustic and visual modalities have different properties, yet humans can seamlessly connect and integrate them to perceive the world. Developing learning algorithms to replicate these abilities, especially for multi-modal audio-visual *fusion* and *retrieval* is of great interest. For this purpose, I built a self-supervised audio-visual model called Contrastive Audio-Visual Masked Autoencoder (CAV-MAE) which was the first model to learn a *joint* (for fusion) and *coordinated* (for retrieval) representation simultaneously [7]. The key insight of CAV-MAE is that a combination of mask data modeling and contrastive objective is crucial for audio-visual model training, which significantly influences subsequent audio-visual model designs, such as Meta's MAViL, that also combine these two objectives. My other research on this direction includes [8]. **CAV-MAE is covered by MIT News[4] and is a notable-top-25% paper of ICLR 2023.**

[1] https://huggingface.co/MIT/ast-finetuned-audioset-10-10-0.4593
[2] https://www.youtube.com/watch?v=uqsW2eK-Rms
[3] https://www.saneworkshop.org/sane2023
[4] https://news.mit.edu/2023/scaling-audio-visual-learning-without-labels-0605

2.4 Secure Audio and Speech AI

With the fast-growing number of users and usage scenarios, security concerns about audio AI (e.g., Amazon Alexa) have become more prominent. My work [9] is one of the first to expose the vulnerability of audio AI systems to *audio adversarial attacks*. That is, audio AI systems, including those for security-critical tasks like speaker verification, can be compromised by very small, yet intricately crafted perturbations of the input signal, resulting in incorrect predictions. To safeguard audio systems from adversarial and other types of attacks, I proposed a defense strategy based on sound source identification to determine whether the source is a loudspeaker (indicative of an attack) or a live speaker (benign). Specifically, I compiled the first *multi-channel* replay attack dataset named ReMASC [10] and developed a neural network-based *multi-channel* sound source detection algorithm [11]. This work has shown that spatial information from multi-channel audio significantly bolsters defense capabilities. **The ReMASC paper received a nomination for the Interspeech 2019 Best Student Paper Award.**

2.5 Speech-Based Health Systems

Speech contains a lot of information including the health condition of the speaker. Compared with other biomarkers, speech is convenient to collect at almost no cost. As my first Ph.D. project, I developed a speech-based depression automatic detection system based on interviews of about 30 minutes [12]. The key novelty of this work is to first logically organize short-term details based on *topic* and then extract *topic*-level acoustic and linguistic features. This is because the same features in different *topics* convey different information and should be regarded as separate features, e.g., smiling in the context of discussing family can be more discriminative than smiling in the context of greeting someone, because the latter might only be due to etiquette. With this technique, **my model won the audio/visual emotion challenge (AVEC) at ACM Multimedia 2017.** Besides this, I have also worked on autism [13], emotion disorder [14], and dementia [15] detection.

3 Future Directions

As the path to Artificial General Intelligence (AGI) unfolds, I aim to leverage my past experiences to develop **Audio and Speech AGI** that is not only smarter but also safer and yields a more substantial positive social impact. My specific interests include the following areas:

3.1 Unified Model for Audio and Speech Perception, Understanding, and Generation

There are many audio and speech tasks such as speech recognition, voice biometrics, audio event detection, and music generation. Currently, each task is typically processed with a *specialized* model. However, biology research suggests that specific regions of the human brain, like the Superior Temporal Gyrus, can process diverse types of auditory information, indicating that the same neurons may handle various audio and speech tasks. In addition, humans can learn a *new* audio or speech task with just a few examples or even from instructions alone. This disparity between existing audio AI and human capabilities motivates my interest in building a **unified audio model** that can solve a wide range of audio and speech tasks from perception to generation, and generalize to new tasks with minimal examples or instructions, akin to GPT's approach to natural language tasks. Such a model should include an audio perception module (an audio encoder), an understanding and reasoning module (a large language model), and an audio/speech generation module (a vocoder or diffusion model). Additionally, novel training data and algorithms are necessary. In advancing computational methodologies, the development of a unified audio model represents a significant step towards the goal of achieving artificial general intelligence, where a machine's cognitive abilities can be generalized across a broad spectrum of tasks similar to human intelligence. I plan to seek funding for this direction through NSF-IIS as well as industry funding opportunities.

3.2 Make Audio and Speech AI Safe, Trustworthy, and Copyright-Respecting

Audio AI, particularly generation models, promise substantial benefits for human productivity and connectivity, yet they also pose a risk of negative social consequences if not developed and managed with caution. For example, modern text-to-speech systems can imitate someone's voice with just a 3-second sample; music generation AI might unintentionally generate pieces that impinge on copyright protections; and multi-modal large language models are prone to hallucination issues, where generated content deviates from factual accuracy.

Addressing the issues of safety, trustworthiness, and copyright respect in audio AI is as critical, if not more so, than enhancing AI's capacity. I am interested in developing secure and ethically sound audio AI systems. Specifically, I aim to develop audio watermarking policies and algorithms that have minimal impact on sound quality while ensuring traceability and permanence; to build defense algorithms that protect audio AI systems against adversarial attacks; to devise systematic evaluation methods for the severity of hallucinations in audio large language models, and to create training and inference algorithms that mitigate such hallucinations; and to construct audio, speech, and music generation algorithms that not only create high-quality outputs but also strictly adhere to copyright rules. I plan to seek funding for this direction through the NSF Secure and Trustworthy Cyberspace (SaTC).

3.3 Bring Speech Health Systems from Lab to Real-World Applications

Health-related research has been a long-standing interest of mine, primarily due to the societal benefits it yields. This starts from my undergraduate research (a smart ECG monitor that won a national award) to my Ph.D. research (several speech-based neuropsychological disorder detection systems), and I am continuing to work in this direction at MIT (in collaboration with Takeda Pharmaceutical Company). Speech, as a biomarker, is notably convenient and cost-effective to collect, often simply through a mobile app, which allows for the development of highly affordable speech-based health systems. Despite the potential, there are currently no large-scale speech-based automatic diagnosis systems in widespread use. The reasons are multi-fold: limited patient speech data restrict the size and the generalization capabilities of models; public datasets may not accurately reflect the demographics and conditions of real-world data; and medical applications demand exceptionally high standards of robustness and explainability.

My ambition is to tackle these issues and transition speech-based health systems from research prototypes to widespread clinical use. Particularly, I plan to form partnerships with clinics and medical institutions to gather speech data from patients in realistic settings, and lower the training data demand with pretraining and transfer learning strategies. Further, I am interested in constructing robust and interpretable models utilizing techniques like Bayesian neural networks, and integrating these models within a *physician-in-the-loop* framework, ensuring they meet practical medical needs. Ultimately, my aim is to innovate speech-based health systems that can transform the management of neuropsychological disorders, fostering prompt diagnosis and enabling early intervention for millions. I plan to seek funding for this direction through NIH with medical professional collaborators.

4 Concluding Thoughts

I am eagerly looking forward to the prospect of taking on a professorial role, where I can collaborate with other exceptional faculty members. I am confident that I will be able to secure external funding to build my research group with diverse and talented students. I will use my expertise to mentor my students to contribute to top-tier publications, and advance audio AI to be smarter and more secure, striving toward Audio and Speech AGI that could revolutionize communication, enhance daily life, and offer substantial societal benefits.

References

[1] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021.

[2] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. SSAST: Self-Supervised Audio Spectrogram Transformer. In *AAAI Conference on Artificial Intelligence*, 2022.

[3] Yuan Gong, Yu-An Chung, and James Glass. PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[4] Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. Whisper-AT: Noise-Robust Automatic Speech Recognizers Are Also Strong General Audio Event Taggers. In *Interspeech*, 2023.

[5] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, Think, and Understand. *arXiv preprint arXiv:2305.10790*, 2023.

[6] Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint Audio and Speech Understanding. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023.

[7] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive Audio-Visual Masked Autoencoder. In *International Conference on Learning Representations (ICLR)*, 2023 (**notable-top-25% paper**).

[8] Yuan Gong, Alexander H Liu, Andrew Rouditchenko, and James Glass. UAVM: Towards Unifying Audio and Visual Models. *IEEE Signal Processing Letters*, 2022.

[9] Yuan Gong, Boyang Li, Christian Poellabauer, and Yiyu Shi. Real-time Adversarial Attacks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

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

[11] Yuan Gong, Jian Yang, and Christian Poellabauer. Detecting replay attacks using multi-channel audio: A neural network-based method. *IEEE Signal Processing Letters*, 2020.

[12] Yuan Gong and Christian Poellabauer. Topic Modeling Based Multi-Modal Depression Detection. In *Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017 (**depression challenge winner**).

[13] Yuan Gong, Hasini Yatawatte, Christian Poellabauer, Sandra Schneider, and Susan Latham. Automatic autism spectrum disorder detection using everyday vocalizations captured by smart devices. In *ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2018.

[14] Yuan Gong and Christian Poellabauer. Continuous Assessment of Children's Emotional States Using Acoustic Analysis. In *IEEE International Conference on Healthcare Informatics*, 2017.

[15] Nauman Dawalatabad, Yuan Gong, Sameer Khurana, Rhoda Au, and James Glass. Detecting Dementia from Long Neuropsychological Interviews. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022.