Statement of Purpose

Jiatong Shi *Ph.D. Applicant*

During my undergraduate studies at Renmin University of China, Professor Qin Jin introduced me to speech processing and instilled in me an interest in this field that remains strong to this day. I subsequently enrolled at Johns Hopkins University to continue my training under Professor Shinji Watanabe, a noted expert in end-to-end automatic speech recognition (ASR). Besides providing me with a strong foundation in computational research methods, the greatest impact that my two advisors have had on me is to encourage me to think across disciplinary boundaries. When perspectives and methods from two different fields are brought together, many exciting possibilities often emerge. At Johns Hopkins, I actively pursued a multidisciplinary approach on three projects in which I participated as a key member of a research team.

- *Target-speaker Speech Recognition*: This project was carried out during my 2020 summer internship at Tencent AI Lab (Seattle), after I had finished one year of the master's degree program at Johns Hopkins University. The Tencent project sought to isolate and transcribe the audio from a targeted speaker by extracting the relevant acoustic signal from noisy environments that included background noise and interfering speakers. The researcher team in which I participated designed a joint framework that combined time-domain target-speaker speech extraction with a Recurrent Neural Network Transducer (RNN-T). A significant challenge to joint training, however, is how to compensate for residual noises and artifacts introduced by the extraction process itself. To deal with this issue, I was responsible for combining neural network-based speaker identification and speech enhancement uncertainty estimation for the joint system. My experiments showed that adding the neural uncertainty module significantly reduced (by 17%) the relative character error rate on multi-speaker signals with background noise compared to the baseline.

  From my participation, I learned how to identify potential problems in a joint system using bad cases. After finishing a joint-training framework, I was hesitant about what my next step would be. My mentors at AI Lab, Dr. Chao Weng, and Dr. Chunlei Zhang, suggested me to go over the cases where our systems could not perform well. From that time, I started to realize residual noises and artifacts introduced from the speech extraction process and finally came up with uncertainty features.

  As I noted below, my work at Tencent has motivated a desire to continue to pursue the problem that noisy environments pose to ASR as these conditions are very common in real life situations and pose a basic challenge to computational tools that work well in more controlled environments.

- *Endangered Language (EL) Documentation*: With Dr. Jonathan D. Amith, an independent scholar who has worked for close to two decades on EL documentation, I launched a project to employ end-to-end ASR to address what has been called the "transcription bottleneck". In EL documentation, the "transcription bottleneck" is the result of a shortage of effective human transcribers for languages that before documentation efforts were often unwritten or at the least lacked speakers skilled in a viable and consistent orthography. ASR has been suggested as a tool to overcome such bottlenecks. In applying ASR to Yoloxóchitl Mixtec, an EL from west-central Mexico with a challenging tonal phonology, we found that the ASR system still made significant errors, particularly in the noisy conditions that often affect EL documentation. To get better results, one approach we pioneered was to initiate transcription with a novice speaker, still with significant deficiencies in transcription, and then use ASR to correct the novice transcription. Our combined or hybrid approach effectively achieved results significantly better than either the novice or the stand-alone ASR transcriptions. This experiment has demonstrated how ASR systems and novice transcribers could work together to improve EL documentation without burdening expert speakers whose time is a scarce resource. We believe this combinatory methodology would help mitigate the transcription bottleneck that hinders EL documentation in general.

  Besides learning how to apply ASR to a minority language, very distinct from the languages most often used in ASR development, this project taught me how to combine linguistic analysis into engineering implementation. After examining the ASR and novice errors, Dr. Amith proposed some heuristic rules for novice transcription correction by noting differences in the types of errors that the novice was most prone to make and those that were more common in ASR hypotheses. The idea was to combine the transcriptions taking advantage of the strengths and accuracies of each. This requires an accurate alignment between ASR hypotheses and novice transcription. However, the simple dynamic programming (DP) for words did not perform well, it was either too general or detailed. Based on the hierarchical structure of different linguistic units, my solution was to introduce character and syllable mismatch to the dynamic function for word-level alignment, achieving 25% relative improvement in transcription (CER) accuracy over the baseline using simple DP.

- *Singing Voice Synthesis (SVS)*: Neural network-based singing voice synthesis systems require sufficient data to train well. Due to high data acquisition and annotation cost, however,data limitation often inhibits effective building of SVS systems. This data scarcity makes neural network models prone to over-fitting. In my work on

this project, I proposed to use a Perceptual Entropy (PE) loss to regularize the network where PE is derived from a psycho-acoustic hearing model, frequently used in speech coding. On the open-source Kiritan database (1 hour), our team explored the impact of the PE loss on various mainstream sequence-to-sequence models, including an RNN-based model, a transformer-based model, and a conformer-based model. Our experiments showed that the PE loss could mitigate the over-fitting problem and significantly improve the synthesized singing quality reflected in objective and subjective evaluations.

Unlike text form, speech acoustics are physical processes. In this project, PE is a physical feature of audio signals. The improvement in objective and subjective evaluation gained from PE loss reminded me that some natural constraints from physical rules might be combined into the algorithms.

The preceding three experiences have led me to recognize two main challenges to the speech processing community. The first is the research challenge of handling multi-party conversations, a problem compounded by other factors that may create a noisy environment. The second is from the engineering side and focuses on boosting research efficiency for the research community by creating reproducible open-source works. I plan to continue working on these two problems during my Ph.D. studies. The following presents an initial effort to meet these two challenges:

- *Natural Multi-Party Conversation Understanding*: Speech in multi-party conversation is often recorded by far-field microphones. Current ASR systems still perform poorly when handling this kind of speech. This is because the signal-to-noise ratio between the target speaker and interfering signals (e.g., background noise, reverberation, and interfering speakers' speech) is much lower than clean, close-talk speech. Processing multi-party conversation speech is a significant problem that needs to be addressed. The speech in real-world conditions often suffers from its effects, such as the famous cocktail party scenarios.

  During my previous work at Tencent AI Lab (Seattle), I was part of a team investigating target-speaker speech recognition using the target speaker's speech as enrollment speech. To implement this approach, however, enrollment speech must be obtained before attempting extraction and ASR of the target speaker. This is often not possible in real-world scenarios. Related tasks of far-field ASR include speaker identification, speech separation and enhancement, and ASR. I am interested in combining these three methodologies into a single system. After simply cascading each module, many research questions will arise in regards to each module's performance. For example, in my recent work, I integrated a speaker diarization system with a target-speaker speech extraction module. The system explicitly obtains speaker embeddings from the multi-talker speech stream and applies the embeddings for speech enhancement and separation. I will be working on this scenario first and then extend the system to recognition.

- *Research Efficiency Elevation*: Since neural network-based methods began to pervade the speech community, there has been a growing concern with reproducibility. I believe that it is important for researchers to share their ways of building models and have been lucky to join the development group of ESPNet, an end-to-end speech processing toolkit. As a member of this group, I have contributed several recipes for multi-lingual ASR and EL documentation. With ESPNet, developers can easily use the state-of-the-art models of several speech tasks (e.g., ASR, speech enhancement, and speech synthesis) and reproduce the experimental results from re-training or pre-trained models.[1]

  Inspired by the open-source experience, I have also launched a singing voice synthesis open-source project, namely "SVS_system", focusing on end-to-end SVS models.[2] Because of high data annotation cost and more strict copyright issues, SVS has suffered from limited open-source data and codes. With the SVS_system I have developed, I hope to offer state-of-the-art solutions for SVS tasks and help support better comparison between models in this field. This project's potential outcome can provide benchmarks for open-source SVS databases and encourage more researchers to study this task.

After three years of studying speech processing with Professors Watanabe and Jin, I am eager to dedicate myself to more focused and meticulous work in speech processing and begin to answer the issues raised in the preceding paragraphs. As physical rules naturally confine speech, I believe research from the engineering side can offer groundbreaking solutions to the challenges we face. After completing my Ph.D. study, I hope to continue to work on interesting academic and industrial projects. If possible, I hope to follow my advisors' paths and inspire the next generation of scholars on new ways of thinking about problems and new approaches to research. Carnegie Mellon University, the alma mater of Professor Jin and the soon-to-be-home of Professor Watanabe, has one of the most talented and diverse faculties for speech processing and NLP. It would be an honor to study at CMU as both of my mentors have eloquently mentioned the cutting-edge approach of the Language Technology Institute faculty and the rigorous nature of their research experiments. I hope I am given the opportunity to continue my research at CMU, with Professor Shinji Watanabe and the other faculty members whose work I have studied.

[1] `https://github.com/espnet/espnet`

[2] `https://github.com/SJTMusicTeam/SVS_system`