Saikrishna Rallabandi web: https://www.cs.cmu.edu/~srallaba srallaba@andrew.cmu.edu

Statement of Purpose

Most depictions of a sufficiently advanced human civilization involve a speech driven interaction between humans and machines and perhaps rightfully so. Speech is the most natural form of communication for humans and engaging conversations are proven stress busters. Over the years, speech technologies have transitioned from working for '*some* people in *some* scenarios' to '*most* people in *most* scenarios'. In the recent years they have also been successful at engaging humans in a limited domain banter. However, speech based interaction with machines or devices has not yet become an indispensable part of our lives and has had minimal impact. This highlights the design problem: How do we design systems that are more intuitive? On the other hand, speech technology has already begun permeating into our lives in non-obvious ways via medical transcriptions, couples' therapies, etc raising an efficiency problem: How do we build systems that are interpretable and have infinitesimally small errors? Problems like this have motivated me to apply for a PhD in Language and Information Technologies.

I am currently a second year MLT student at LTI, CMU advised by Prof. Alan W. Black and before that, a Masters student at IIIT Hyderabad advised by Dr. Kishore Prahallad. Both these mentors have taught me to always try and answer the following questions as a research student: Can we make fundamental observations based on the data and exploit its inherent structure to design efficient techniques? Can we bring human into the loop and apply specific cues that help algorithms achieve high rates of acceptance? I strongly believe that this approach of thinking leads to techniques that have longevity. I will describe three of my projects highlighting this style of thinking:

- *IIIT-H System for Blizzard Challenges 2015 and 16[1] :* I have built a full scale independent synthesis system as a part of our submission to Blizzard speech synthesis challenge 2015 and it was the first time we participated. The system was a syllable based unit selection and concatenation one. It handled six Indian languages using reduced vowel based Epenthesis for missing syllables and word to phone mapping for English words. It is interesting to note that both Epenthesis and word to phone mapping are linguistic theories that explain human behavior in similar situations. To be more specific, when faced with the issue of having to pronounce a consonant cluster not supported by the phonotactics of their native language, humans tend to introduce an extra vowel between the two consonants. When pronouncing a foreign word, speakers are known to first form a mental mapping in their native language. Based on these insights, I have developed algorithms that mimic such behaviors into our submission. The results clearly indicate the contribution of these techniques to MOS scores of our system (code E), especially in the Multilingual task where our submission significantly outperformed other competitors. In addition to teaching me essential system building skills, this project has helped me apply well studied linguistic theories which provide practical insights to solve practical problems.

- *Audio Rendering of STEM Content:* In this project, I have developed a screen reader for students with print disabilities. Based on the observation that human mentors modulate their vocal characteristics while training such students, I have designed 5 specific prosodic manipulations to a prebuilt voice helping the comprehensibility of the content. For instance, one variant employs pitch variation to help demarcate the base term from exponent while rendering a mathematical equation. The system is currently deployed to aid the visually challenged people at LVPEI Hospital, Hyderabad enabling them to prepare for competitive bank exams. This experience has taught me the significance of having human in the loop while attempting to provide solutions especially in challenging scenarios.

- *Submission from CMU to Voice Conversion Challenge 2018:* I have built from scratch our entry to Voice Conversion Challenge 2018. The constraint of the challenge was the ridiculously small amount of data, 5 minutes per speaker. For comparison, we typically need at least half hour of speech to build a reliable speech synthesis system. As I was the only student working on the project, it was paramount to prioritize the ideas being experimented on. To accomplish this, I have preselected modules in the pipeline based on a rank ordered list of return expectation, dismantled all known intuitions and exhaustively experimented every possible explanation in the configuration space. This has helped equip our submission with novel alignment and modeling algorithms that have potential to push the current boundaries in the domain of voice conversion. I am hopeful that the built modeling infrastructure will

replace its counterpart - and the strongest component - in current speech generation pipeline of CMU. Working on this project has taught me the the significance of questioning the current practices and then validating hypotheses using extensive experimentation. More importantly, this exercise made me incorporate practical hacks such as buffer times and aggressive scheduling to the way I manage time and work load.

The experience gained from all these projects leads me to believe that we need to solve two challenges before speech technology becomes ubiquitous: the research challenge of developing strategies to handle new phenomenon in conversational speech such as code mixing and the engineering challenge of encapsulating current research so that developers can build applications on top. I am motivated to work on these two problems during my PhD and will describe the work done so far.

- *Code Mixing:* Code Mixing or Code Switching is a phenomenon where multilingual speakers tend to use more than one language in a single sentence. Typical speech processing systems are built to handle a single language and tend to ignore the contents of the other languages while processing such mixed data. However, Code Mixing is pretty common in multilingual societies like India and Singapore to an extent that most conversations and newspaper headlines are mixed. Hence, it is important that we work on techniques which can handle mixing. To render mixed content using Text to Speech, I have been working on strategies at every level[2,3,4] in the voice building process. Consequently, I was able to come up with a deceptively trivial implementation using signal processing and latent stochastic models that can now enable us to adapt an existing speech synthesis system to any desired language. I am ultimately interested in designing a controllable system. This means we can modify a parameter akin to a knob and generate disparate accents, styles and speakers using a single voice. Simultaneously I plan to solve the identification counterpart in two phases where the first phase deals with detecting such content followed by a recognition phase. I have built a functional near cent accuracy detection system and I will be working on the recognition phase with an intention to delineate and minimize the errors made by current ASR systems.

- *Open Advancement:* Speech research has been dominated by complex systems that require significant expertise. This prevents non experts from integrating existing research into applications. I believe that it is important for researchers to package and expose their work (as APIs) so that it becomes effortless for the developers to extend ongoing research and build exciting real world applications. I have been working with Prof. Alan Black for over a year on Flite - a run time speech synthesis engine from Festival. With Flite, our goal is to generate a single file which developers can use to integrate into their applications. We have demonstrated the process by deploying Flite onto Android and Windows platforms. Simultaneously, I have been working with Prof. Eric Nyberg on an early phase long term project with an end objective of Open Advancement (OA) for speech processing. With OA, I wish to make a transition from the practice of sharing code to sharing services on cloud based containers that enable rapid prototyping and experimentation. This means the developers can not only just integrate final research outcomes into their applications but also experiment with the existing research architectures - all without having to comprehensively learn speech processing.

Having spent two years at IIIT-H followed by two years at LTI, I now want to make a far meticulous and focused attempt at research in speech and NLP, one with an attempt to answer the questions mentioned in this statement. I believe that the developments in AI are projecting these domains from low stake - low risk to high stake - high risk category. Therefore, for reasons beyond the objective of achieving the best metrics and outclassing the competitors, it is important to work on problems that have interpretibility, push the boundaries of the domains and result in viable, practical applications. In general, I am convinced that core engineering research can provide elegant solutions to all the challenges we face as humanity today. I envision myself working on audacious academic and industrial projects after PhD and hopefully inspire the next generation of students to consider research as a possible career path. Carnegie Mellon University has always been aggressively biased towards ambitious projects with overarching goals. In addition, I have experienced first hand the way Professors here go out of their way to provide the required resources and give enough space and time for students to grow into able researchers. I therefore wish to continue my work here, specifically with Prof. Alan who has been a torch bearer of innovation in my domain.

1. http://researchweb.iiit.ac.in/~saikrishna.r/Blizzard_2015_submission
2. Experiments with Cross-lingual Systems for Synthesis of Code-Mixed Text, SSW9, 2016.
3. On building mixed lingual speech synthesis systems, Interspeech 2017
4. Speech Synthesis for Mixed-Language Navigation Instructions, Interspeech 2017.