Statement of Purpose

On listening that I work in the area of speech sciences, a technology enthusiast responded curtly saying that he is equally sad today with his gadgets such as mobile phone, etc. as he was ten years before, the only difference being that he can now speak to these gadgets. This comment clearly summarizes the shortcomings associated with my field and the research efforts being carried out worldwide, which if we happen to put as a couple of questions, come out as: How to make the speech technology closer to human use? Can we come up with practical implementations of theoretical aspects of speech and language in order to solve the existing problems? Through my research, I want to address these issues and more importantly, offer simple and elegant solutions to such problems with the help of subtle obervations.

Research

I was introduced to this amazing field of speech technology about eight years ago, during my undergraduation project on Automatic Recognition of Emotions from Speech for which, I have built a system capable of recognizing 4 emotions(anger, sadness, happy and neutral) in real time. This project was instrumental in providing me the basic understanding and background required for processing speech data, such as the block processing. Right after graduation, I have been employed at Amazon Development Centre, India as a digital data analyst and subsequently at Google, India as Maps data analyst, for two years, both positions helping me improve my system building skills.

My real interaction with speech systems research took place when I joined IIIT Hyderabad as a full time research student and my advisor **Dr. Kishore Prahallad** introduced me to the challenging field of speech synthesis, where even the minute glitch matters and a contribution can have tremendous impact. My primary goal as a research student was to address the issues related to Text to Speech in Indian languages and develop systems which are both robust and reliable on the one hand as well as unrestricted in nature on the other, and deploy them on multiple platforms.

As a part of research, I have always tried to answer the questions: Can we make fundamental observations based on the data and exploit the inherent structure to design techniques which improve the performance of the existing system as a whole? Can we bring human into the loop thereby ensuring that we end up with a system that fares well in terms of acceptance in addition to efficacy? I believe that this approach promises good results and also leads to techniques that have longevity. I will describe two of my projects to highlight my application of the aforementioned approach.

IIIT-H System for Blizzard Challenge 2015

Blizzard Challenge is an initiative conducted every year with an aim of analysing best practises for robust speech synthesis. I have built the IIIT-Hyderabad synthesis system[1] for Blizzard 2015, which was the first time we participated in the challenge. The system in a nutshell was a *syllable based unit selection and concatenation* one for six Indian languages with reduced vowel based *epenthesis* for handling the missing syllables and *word to phone mapping* for handling the English words. Both epenthesis and word to phone mapping are essentially techniques designed by making observations of human behavior in similar contexts and then trying to impart the inferences from these observations to the synthesis system.

Humans when faced with the issue of having to pronounce the consonant clusters not suported by the phonotactics of their native language, tend to introduce an extra vowel between the two consoanants to aid the pronunciation. For instance, 'Bulb' is pronounced as 'BalUbU' by a native Telugu speaker, as 'lb' cluster is not supported by the phonotactics of Telugu. Similarly, when required to pronounce clusters missing from their native language, they tend to make a mental mapping from the *word* required to be pronounced to the phones of their native language as opposed to phone to phone mapping used today. The idea was to integrate this behavior into the speech synthesis system and make the system robust. From the results of the challenge, it can be observed that there was clearly a major contribution of these techniques to the improvement in the MOS scores of our system (code E), especially in the Multilingual task in which both of these techniques were deployed extensively.

1. http://researchweb.iiit.ac.in/~saikrishna.r/Blizzard_2015_submission

Audio Rendering of STEM Content

This is an application oriented research project[2] that I was part of where we built synthesis systems capable of rendering STEM content in audio with enhanced prosody, with an intention to design efficient screen readers for students with print disabilities. This was achieved by making specific modifications to the prosody of the prebuilt voice by making observations from human speakers trained to teach such people. We have used the presentation markup of MathML to represent the equations and then designed four systems, each of which made a specific prosodic manipulation and further built a hybrid system integrating the best of the techniques following a subjective evaluation.

The system is currently deployed to aid the visually challenged people at LVPEI Hospital, Hyderabad enabling them as of today to solve complex mathematical equations on par with the normal sighted people. Later, I have created a parser linking the techniques to the SABLE markup language used by Festival Speech synthesis system and made it opensource so that the techniques can used by a wider audience.

In the field of speech synthesis, apart from the mentioned projects, I have developed a continuous representation of input text for better acoustic and prosodic modeling using Positive Pointwise Mutual Information Matrix(PPMI) Factorization approach, by observing that it is analogous to the word2vec model in its functionality and much simpler in terms of implementation. This was extended to model the phone durations from diverse sources such as audiobooks, lecture series and children′s stories to aid the synthesis. The derived representations were also used to model and extract *humor* from social network data. I have also extended the Indian language Unit selection system submitted to Blizzard Challenge by proposing cross fade mechanism considering waveform similarity and later incorporated it as join cost function during the Viterbi search.

Other research projects related to Speech Technology

During Google Summer of Code 2015[3], I have implemented a deep learning based speech recognition wrapper toolkit based on Kaldi for RedHen Lab, whose audio database is a collection of 250,000+ hours of television news in 6 European languages(French, German, Spanish, Norwegian, Danish and Swedish) apart from English. The acoustic models for the different languages are available upon request by the RedHen Lab. Inspired by this progress, I have worked in collaboration with a fellow research student where we developed a speech synthesis system using Kaldi as the front end for text transcriptions[4]. We extended the system further by using the phone confidence level both as a measure of the transcription accuracy and also for data pruning, resulting in an improvement in the speed of the voice building process.

My other research contributions include developing novel authentication systems based on biosignals such as fingersnaps and whistles, whisper speech modeling and anonymization of speakers for privacy protection in Oral personal histories[5], which shall be extended in the ongoing Voice Conversion Challenge 2016, where I will be designing the IIIT-H system.

Conclusion

Having spent some time in industry and then two years at IIIT-H for my Masters, I am always open to collaborative research and as a part of the same, I work in teams and maintain three joint blogs(links in homepage) with my fellow researchers where we post technical content related to speech and also make several podcasts about implementations and nuances of speech systems which are publicly available. That being said, I now want to pursue a Ph.D., as it would allow me to do a much more focused research having a more profound impact on the field and moe importantly, people for their daily work.

At CMU, Prof. Alan Black and Prof. Alex Rudnicky are working on the research problems related to different applications of speech technology and my work is closely related to their research, infact some of them being parallel approaches (Ex. Synthesis for Orthography less languages by Prof. Alan vs Kaldi based synthesis system, Random Forests vs DNN for acoustic modeling in synthesis). Other work I want to mention here is that of speaker anonymization, on which Prof.Bhiksha Raj has worked.

Such parallel efforts have reassured me about importance of the problems I am working on and made me very optimistic that my long term goals can be realized at CMU. I would love to work with any of the faculty at LTI and be able to contribute positively to the research in the lab.

2. http://researchweb.iiit.ac.in/~saikrishna.r/Audio_Rendering_of_STEM
3. http://www.redhenlab.org/gsoc2015/rsoc15report
4. http://researchweb.iiit.ac.in/~saikrishna.r/Speaker_Anonymization
5. http://researchweb.iiit.ac.in/~saikrishna.r/Kaldi_based_Synthesis