Thesis: Identifying, Evaluating, and Applying Importance Maps for Speech

Abstract: Like many machine learning systems, speech models often perform well when employed on data in the same domain as their training data. However, when the inference is on out-of-domain data, performance suffers. With a fast-growing number of applications of speech models in healthcare, education, automotive, automation, etc., it is essential to ensure that speech models can generalize to out-of-domain data, especially to noisy environments in real-world scenarios. In contrast, human listeners are quite robust to noisy environments. Thus, a thorough understanding of the differences between human listeners and speech models is urgently required to enhance speech model performance in noise. These differences exist presumably because the speech model does not use the same information as humans for recognizing the speech. A possible solution is encouraging the speech model to attend to the same time-frequency regions as human listeners. In this way, speech model generalization in noise may be improved.

We define those time-frequency regions that humans or machines focus on to recognize the speech as importance maps (IMs). In this research, first, we investigate how to identify speech importance maps. Second, we compare human and machine importance maps to understand how they differ and how the speech model can learn from humans to improve its performance in noise. Third, we develop a structured saliency benchmark (SSBM), a metric for evaluating IMs. Finally, we propose a new application of IMs as data augmentation for speech models, enhancing their performance and enabling them to better generalize to out-of-domain noise.

Overall, our work demonstrates that we can improve speech models and achieve out-of-domain generalization to different noise environments with importance maps. In the future, we will expand our work with large-scale speech models and deploy different methods to identify IMs and use them to augment the speech data, such as those based on human responses. We can also extend the technique to computer vision tasks, such as image recognition by predicting importance maps for images and use IMs to enhance model performance to out-of-domain data.

Committee:

- Professor Michael Mandel, Mentor, Brooklyn College
- Professor Rivka Levitan, Brooklyn College
- Professor Alla Rozovskaya, Queens College

Outside Member:

- Professor Brian Mcfee, Outside Member, New York University