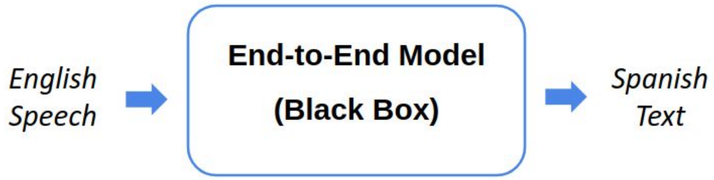
Combining ASR and NMT as *speech2texto*

Nick Hinke and Katie Brandegee

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

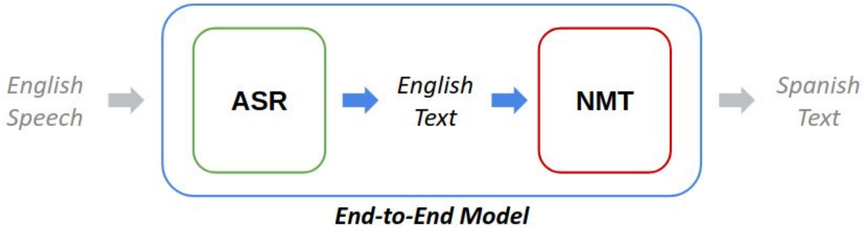
Problem Statement

- **Project Goal:** Convert audio samples of natural language (English) into translated text (Spanish)



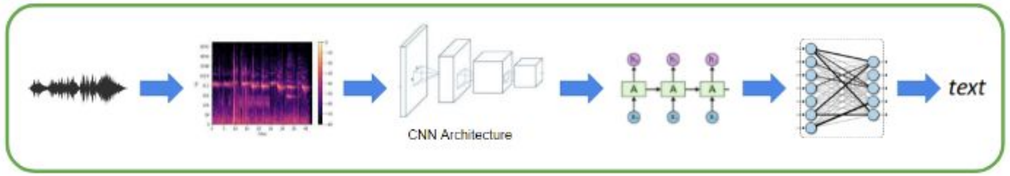
Overall Approach

- **Project Approach:** Construct end-to-end model using two-stage pipeline where each stage can be individually designed, trained, and evaluated before chaining them together
  - *Stage 1:* Automatic Speech Recognition (ASR)
  - *Stage 2:* Neural Machine Translation (NMT)

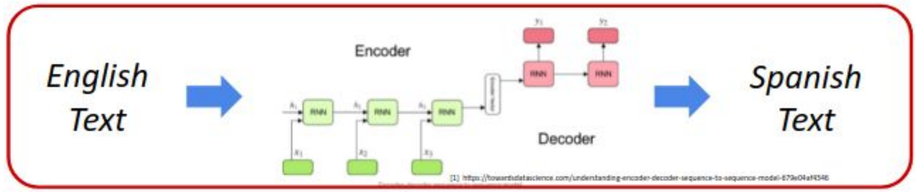


Solution

- **ASR Stage:** Audio Signal --> Spectrogram --> CNN --> RNN --> FC with Softmax --> Text



- **NMT Stage:** English Text --> Sequence to Sequence Model --> Spanish Text



Milestones

- **Preliminary Project Calendar:**

3/20 - 3/26	3/27 - 4/2	4/3 - 4/9	4/10 - 4/16	4/17 - 4/23	4/24 - 4/30
Preliminary Planning and Research	Convert input audio to images, develop an optimal CNN architecture	Develop an optimal RNN architecture, establish a smooth connection between networks, and tune hyperparameters	Develop the sequence to sequence model used in the NMT stage	Train and tune whole model using cross-validation	Address any issues that may arise from language context issues

Proposal Presentation Script

Good morning! I hope you all are doing well. My name is _____ and my partner is _____, and we are both currently in our last semester of the 5 year Robotics BS/MSE program here at Hopkins. Between the two of us, we are involved in quite a variety of subfields within the realm of robotics, all of which have a variety of potential applications for deep learning. As a result, we actually had a pretty tough time determining what we wanted to do for our final project for this course. Ultimately, we decided that despite our lack of experience in the field, we both share an interest in Natural Language Processing, and in travelling the world! So, to that end, we decided to pursue a project that would involve both of those interests.

At a high level, our plan for this project is to construct an end-to-end model that is capable of receiving inputs in the form of English audio signals, and subsequently transform those inputs into transcribed text in a different language. For now, our plan is to first use Spanish as the second language, and potentially add more languages if time permits. That being said, if we were to treat this end-to-end model as a single black box network like it is pictured in the top left, we would more than likely achieve very poor results. Instead, we will construct our model within the framework of a two stage pipeline, where the first stage will convert the English audio signal to English text using the technique of Automatic Speech Recognition, and the second stage will convert the English text into translated text using the technique of Neural Machine Translation. Since both of these processes are already pretty well-defined, this pipeline framework will greatly increase our chances of success, and will also allow us to individually evaluate the performance of each stage of the model.

Obviously, determining the exact architecture within each stage will require a good bit of experimentation, but we have at least developed an initial plan to get started. Just to quickly highlight each stage, the ASR stage will first convert the audio signal into a mel spectrogram which can be understood as an image by a CNN. The CNN will then extract features from the audio image and pass them along to a RNN which will incorporate the temporal information in the data. Then finally, the outputs of the RNN will be fed through a few fully connected layers to produce characters and words which will hopefully be comprehensible. Then in the NMT stage, the english text outputs from the ASR stage will first be converted into viable network inputs using one hot encoding before being fed into a sequence to sequence model based on RNNs. This model should then be able to extract a lot of useful sequential information from the data, and finally spit out a translated version of the input text.

Lastly, as you can see on the bottom right, we have set some soft preliminary “deadlines” to keep ourselves on track. Our only real concern here is that we may be getting a bit too ambitious with this project. We really do believe that we can accomplish this task, but we are a little worried that the complexity of this project may exceed the time we have available to complete it. Thankfully though, if we do find ourselves overwhelmed, we can focus on just one stage of the pipeline, which would likely be the first stage of automatic speech recognition. On the other hand, if the opposite is true and we find that this project wasn’t challenging enough (however unlikely that may be), we could always add a third stage to our framework that involves synthesizing the translated text back into an audio signal. What’s more, given the inherent complexity of our model, we will necessarily require a substantial amount of training data. Fortunately though, this should not be a concern due to the plethora of data available within publicly available datasets from Mozilla and the Tatoeba Project. But regardless of what exactly the final architecture looks like and the data that that goes into it, we’re really excited about this project, and we are really just looking forward to getting started! Thanks so much.