

## **Summaries of Teammates's Reports**

### **#1 Amanda's Report - Deep Neural Network Architectures for Speech Recognition**

- Recent developments in the Speech Recognition focus on applications of deep neural networks, moving away from traditional approaches like Gaussian Mixture Models and Hidden Markov Chains.
- Current research is focused on identifying the most optimal representations for the speech features, which can be used to obtain the best results.
- SincNet Architecture- Uses raw audio frames as input allows "the networks to "learn low-level speech representations of the waveforms", as opposed to networks that take input in the form of human-crafted features (such as Mel Frequency Cepstral Coefficients)."
- No need for human feature design, saving time.
- "Certain valuable traits, such as fundamental pitch and formant frequency information of the signal are maintained". The approach minimizes the number of parameters while preserving signal information, reducing compute complexity.
- By reducing learning complexity, the model becomes very efficient at learning features, leading to lower training times.
- Another approach by Seki et al involves a bank of Gaussian Filters, the approach being somewhat similar to the SincNet approach, reducing compute complexity by learning fewer parameters.
- 3D CNNs: This approach utilizes a 3-dimensional CNN, and uses MFEC inputs, instead of raw audio frames. The goal is to create a speaker model that is independent of the speaker setting, therefore accurately recognizing a word utterance irrespective of its intonation and other characteristics being somewhat different.
- While there is no consensus on any one model being the most optimal, deep learning approaches are currently state of the art. They are however computationally very intensive and may be accessible to only well-funded labs and teams.

### **#2 Shineun's Report - Deep Neural Network Architectures for Speech Recognition**

- "Digital Speech Processing is a broad field of study that embraces: speech recognition, speech synthesis, speaker recognition, language identification, lip synchronization, and co-channel separation."
- Evolution of Algorithms: Kalman filters were the first algorithms used in digital speech processing. They consist of two steps, first determining the values for state variables, along with their uncertainties. In the next step, the algorithm updates these values using a weighted average, with more weight being assigned to estimates with higher certainty.
- Multi Speaker algorithms involve "multi-speaker detection, tracking, and segmentation of speakers".
- Problems faced by Kalman filters include speaker overlap and break and silences in the signal.

- To overcome these problems, Hidden Markov Models proposed which utilize parallel Kalman filters.
- APIs: Google speech to text API, Microsoft Speech Service API: Text-to-Speech, IBM Watson's Speaker Speech-to-Text API are a few public APIs for speech processing.
- The Google API implements an RNN to model speaker embeddings.
- It can recognize 120 different languages but is cost-prohibitive.
- The Microsoft Speech service API can recognize individual speakers in a conversation and is useful for the analysis of a multi-speaker environment.
- However individual speaker samples are required for each speaker in the environment and there are various format specifications that need to be conformed to.
- Each API is better suited to specific applications, with the Microsoft API more useful to enterprise users and the Google cloud API better suited for non-commercial applications.