Machine Learning in Speech Processing: Speaker Diarization

Nihar Dwivedi
ndwivedi@bu.edu
Boston, MA

Introduction

Speech processing has been an important research area of computer science and signal processing for the last few decades. Recent advances in machine learning have had a profound impact on the field. Long gone are the days in which automatic transliteration used to be riddled with errors and accuracy problems, and you needed to manually train the model. These days, the latest version of Google Assistant can recognize speech on the device locally and instantaneously, without any internet connection. Cloud platforms even offer API endpoints for speech to text conversion, speaker identification, and dialogue labeling. It seems a solved problem. But is it? There are still plenty of challenges left to solve, like real-time multi-language translation, on-device translation, etc.

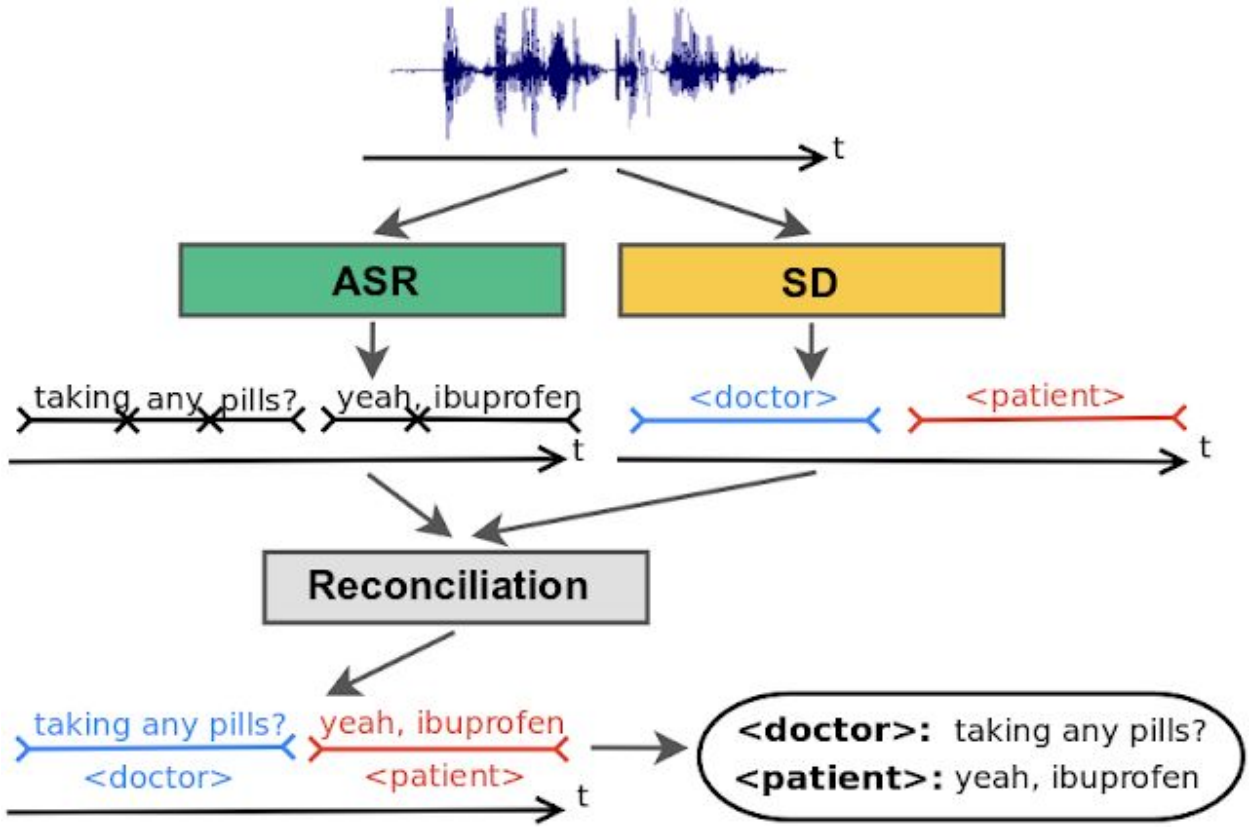
Analysis

In their seminal paper titled "Joint Speech Recognition and Speaker Diarization via Sequence Transduction", the Google AI teams have demonstrated the suitability of a new neural network model, Recurrent Neural Network Transducer(RNN-T) to speaker diarization. Using this model, the team achieved a new performance breakthrough, from 20% to 2% in word diarization error rate, a factor of 10 improvement.

Conventional Speaker Diarization Systems
Ordinary speaker diarization frameworks depend on contrasts in how individuals sound acoustically to recognize the speakers in a sound sample. While male and female speakers can be recognized effectively from their pitch utilizing basic

acoustic models (e.g., Gaussian blend models) in a solitary stage, speaker diarization frameworks utilize a multi-organize way to deal with speakers having a conceivably comparable pitch. First, a switch model separates the discussion into homogeneous fragments, ideally containing just a solitary speaker, in view of recognized vocal attributes. At that point, deep learning models are utilized to guide sections from every speaker to an inserting vector. Lastly, in a bunching stage, these embeddings are gathered to monitor a similar speaker over the discussion.

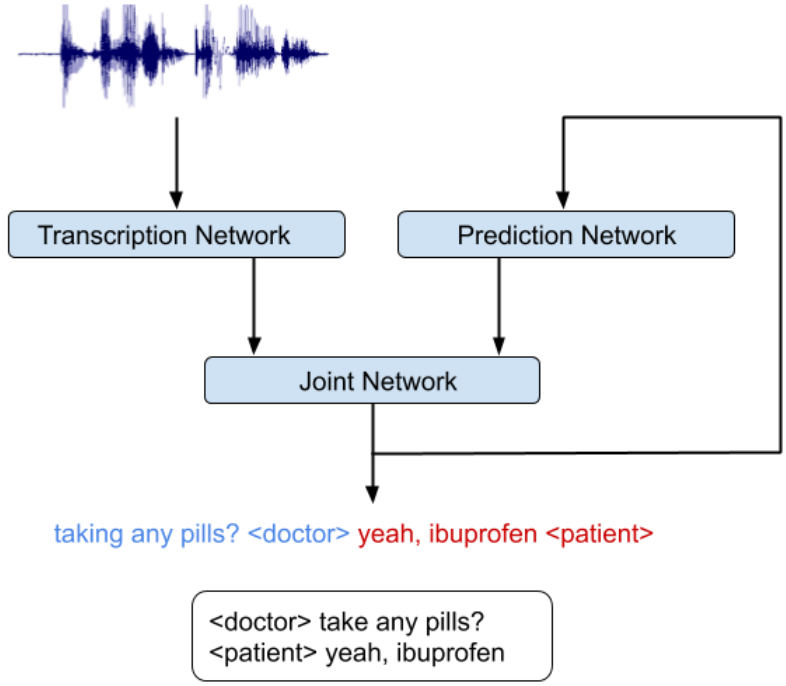
Mostly, the speaker diarization framework keeps running in parallel to the programmed discourse acknowledgment (ASR) framework and the yields of the two frameworks are consolidated to credit speaker names to the perceived words.



Conventional speaker diarization system infers speaker labels in the acoustic domain and then overlays the speaker labels on the words generated by a separate ASR system

An Integrated Speech Recognition and Speaker Diarization System
The group has built a novel and straightforward model that joins acoustic and semantic signs consistently as well as consolidates speaker diarization and discourse acknowledgment into a single framework.

The key knowledge in the paper was to perceive that the RNN-T design is appropriate to incorporate acoustic and phonetic prompts. The RNN-T model comprises of three unique systems: (1) an interpretation organize (or encoder) that maps the acoustic casings to a dormant portrayal, (2) a forecast system that predicts the following objective mark given the past objective names, and (3) a joint system that consolidates the yield of the past two systems and produces a likelihood appropriation over the arrangement of yield names around them.

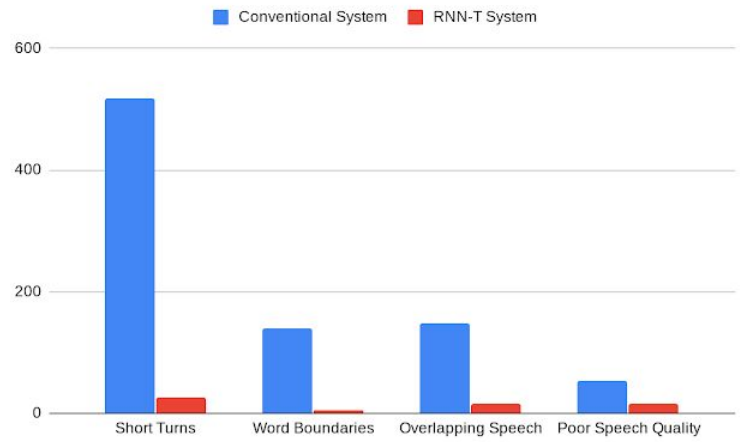


An integrated speech recognition and speaker diarization system where the system jointly infers who spoke when and what.

Training the RNN-T model on accelerators like graphical processing units (GPU) or tensor processing units (TPU) is complex as computation of the loss function requires running the forward-backward algorithm, which includes all possible alignments of the input and the output sequences. This can be resolved by recasting the problem as a sequence of matrix multiplications. The team also took advantage

of an efficient implementation of the RNN-T loss in TensorFlow that allowed quick iterations of model development and trained a very deep network.

The integrated model can be trained just like a speech recognition system. The reference transcripts for training contain words spoken by a speaker followed by a tag that defines the role of the speaker. For example, “When is the meeting?” <CEO>, “It’s at 3 pm,” <Secretary>. Once the model is trained with examples of audio and corresponding reference transcripts, a user can feed in the recording of the conversation and expect to see an output in a similar form. The analyses in the paper show that improvements from the RNN-T system impact all categories of errors, including short speaker turns, splitting at the word boundaries, incorrect speaker assignment in the presence of overlapping speech, and poor audio quality. Moreover, the RNN-T system exhibited consistent performance across conversation with substantially lower variance in average error rate per conversation compared to the conventional system.



A comparison of errors committed by the conventional system vs. the RNN-T system, as categorized by human annotators.

Pros-

This is a state of the art system with record-setting scores. The accuracy of this model is very high, with low error rates.

Cons-

This is a computationally intensive model. The training of this model requires clusters of high-end GPUs or TPUs.

Recommendations

For getting a basic hang of how does speaker diarization work, follow the instructions to use the google cloud speech to text API at the website in the reference [5].

Conclusion

The pace of improvement in speech recognition has been astonishing to see as a result of applied machine learning. The accuracy will surely improve even further, perhaps even surpassing native human capabilities, in the future.

References

[1] Joint Speech Recognition and Speaker Diarization via Sequence Transduction
https://ai.googleblog.com/2019/08/joint-speech-recognition-and-speaker.html

[2] Speaker Diarization
https://en.wikipedia.org/wiki/Speaker_diarisation

[3] Accurate Online Speaker Diarization with Supervised Learning
https://ai.googleblog.com/2018/11/accurate-online-speaker-diarization.html

[4] An All-Neural On-Device Speech Recognizer
https://ai.googleblog.com/2019/03/an-all-neural-on-device-speech.html

[5] Separating different speakers in an audio recording
https://cloud.google.com/speech-to-text/docs/multiple-voices

[6] Alvin F Martin, Mark A. Przybock
Speaker Recognition in a Multi-Speaker Environment
http://www.imm.dtu.dk/~lfen/Speaker%20Recognition%20in%20a%20Multi-Speaker%20Environment.pdf