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.

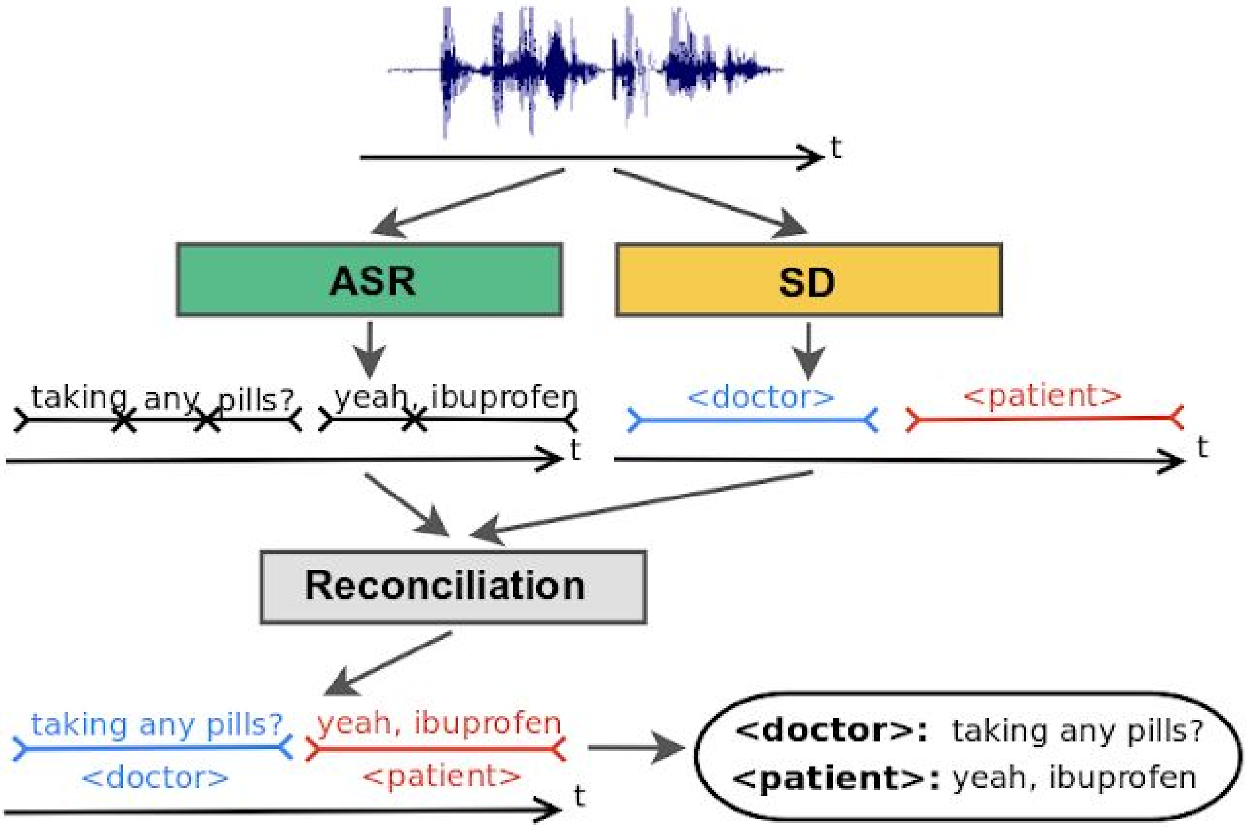
Analysis

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

Conventional speaker diarization systems rely on differences in how people sound acoustically to distinguish the speakers in the conversations. While male and female speakers can be identified relatively easily from their pitch using simple acoustic models (e.g., Gaussian mixture models) in a single stage, speaker diarization systems use a multi-stage approach to distinguish between speakers having a potentially similar pitch. First, a change detection algorithm breaks up the conversation into homogeneous segments, hopefully containing only a single speaker, based upon detected vocal characteristics. Then, deep learning models are employed to map segments from each speaker to an embedding vector. Finally, in a clustering stage, these embeddings are grouped together to keep track of the same speaker across the conversation.

In practice, the speaker diarization system runs in parallel to the automatic speech recognition (ASR) system and the outputs of the two systems are combined to attribute speaker labels to the recognized words.

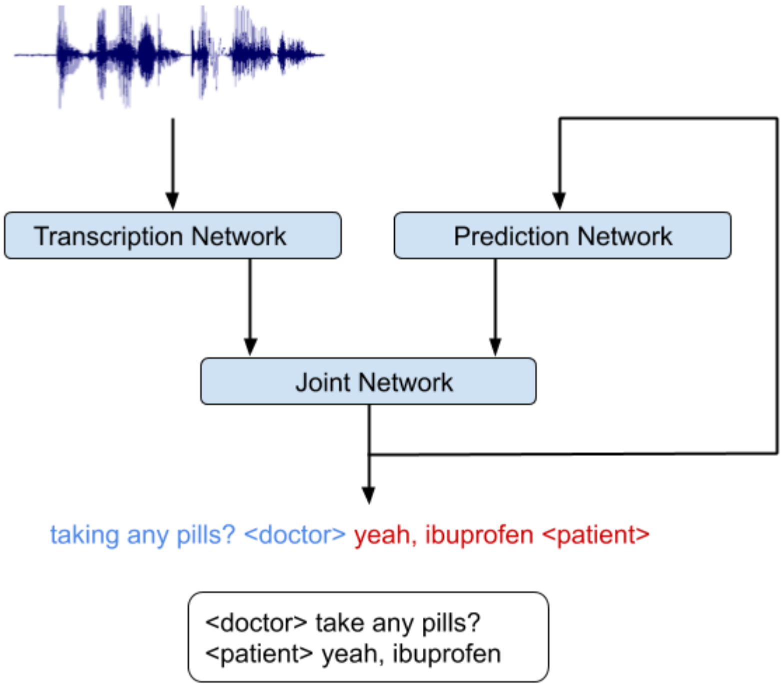


Conventional speaker diarization system infers speaker labels in the acoustic domain and then overlays the speaker labels on the words generated by a separate ASR system

An Integrated Speech Recognition and Speaker Diarization System

The team has developed a novel and simple model that not only combines acoustic and linguistic cues seamlessly but also combines speaker diarization and speech recognition into one system. The integrated model does not degrade the speech recognition performance significantly compared to an equivalent recognition only system.

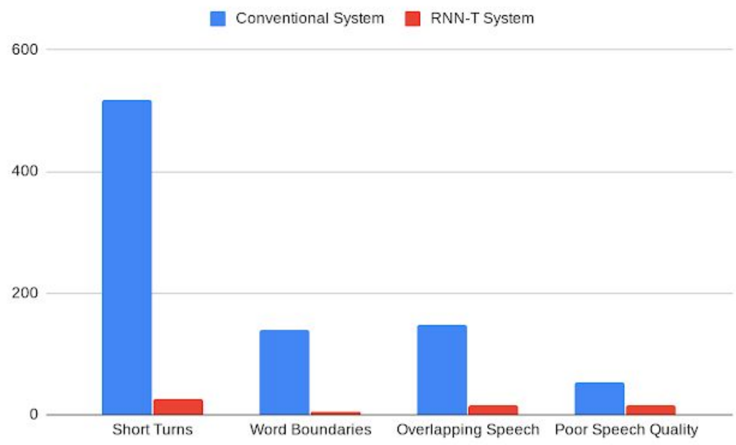
The key insight in the paper was to recognize that the RNN-T architecture is well-suited to integrate acoustic and linguistic cues. The RNN-T model consists of three different networks: (1) a transcription network (or encoder) that maps the acoustic frames to a latent representation, (2) a prediction network that predicts the next target label given the previous target labels, and (3) a joint network that combines the output of the previous two networks and generates a probability distribution over the set of output labels at that time step. Note, there is a feedback loop in the architecture (diagram below) where previously recognized words are fed back as input, and this allows the RNN-T model to incorporate linguistic cues, such as the end of a question.


Transcription Network
Prediction Network
Joint Network
taking any pills? <doctor> yeah, ibuprofen <patient>
<doctor> take any pills?
<patient> yeah, ibuprofen

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 non-trivial 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 issue was addressed recently in a TPU friendly implementation of the forward-backward algorithm, which recasts 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 homework due?” <student>, “I expect you to turn them in tomorrow before class,” <teacher>. 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.

Furthermore, this integrated model can predict other labels necessary for generating more reader-friendly ASR transcripts. For example, the team has been able to successfully improve our transcripts with punctuation and capitalization symbols using the appropriately matched training data. Our outputs have lower punctuation and capitalization errors than our previous models that were separately trained and added as a post-processing step after ASR.

Pros-

This is a state of the art system with record-setting scores. The accuracy of this model is nearly perfect, with very low errors.

Cons-

This is a computationally intensive model. The training of this model requires 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