How to Build and Curate Great Datasets for AI Speech Model Training

Presenter Name
Matthew Lightman

matthew@deepgram.com

IEEE RTC Conference 10/14/2021

Requirements for a Speech Dataset

What features are needed in a speech dataset in order to train AI speech models (e.g. ASR)?

1. Diversity of data
2. Consistency in style of text
3. Ability to filter bad data
4. Ability to filter out-of-domain data
5. Feedback loop with transcription team

Diversity of Data

Difficulty

- Easy (short sentences, simple vocabulary)
- Medium
- Hard (long recordings, complex and idiomatic vocabulary, background noise, cross-talk)

Domain

- Podcast/interview
- Phone Calls
- Meetings
- ...

Languages

- Different training set for different languages
- Dialects
- Accents

Diversity of Data

Data Sources

- Web
- Customers
- Self-Generated: Hire people to create audio content, either from pre-written scripts, or in a more organic setting

Transcription

- Some data transcribed by an in-house transcription team
- Some audio with transcripts come from “the wild”
- With a transcription team, get higher quality transcripts, but may not scale as well as data from other sources.

Consistency in Style of Text

Inconsistencies will confuse models and directly correlate to mis-transcriptions.

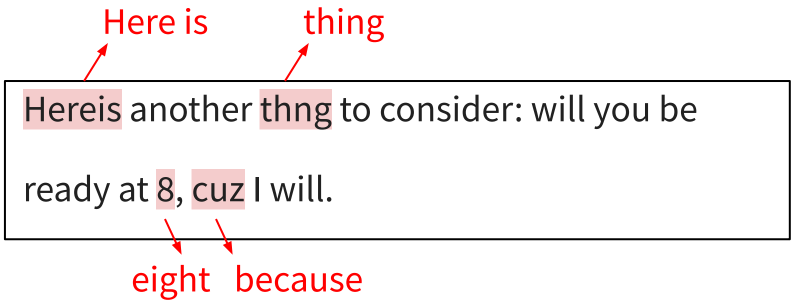
“Ok”, “okay”, and “okayyy” are the same word. Are “k” and “mkay” the same word too?

How to represent numbers?
“12” → “twelve”
“2nd” → “second”

Consistency in Style of Text

- Transcriptionists have style guides that are developed in collaboration with data scientists.

- “Text cleaning” to standardize transcripts that come from other sources than the transcriptionists. Uses a spellcheck-like methodology, in addition to curated rules and regular expressions.



Here is thing

Hereis another thng to consider: will you be ready at 8, cuz I will.

eight because

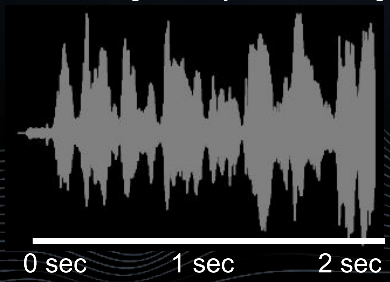
Filtering Bad Data

- Don’t delete data in case it’s useful in the future. Apply filters that are customized to the particular task.

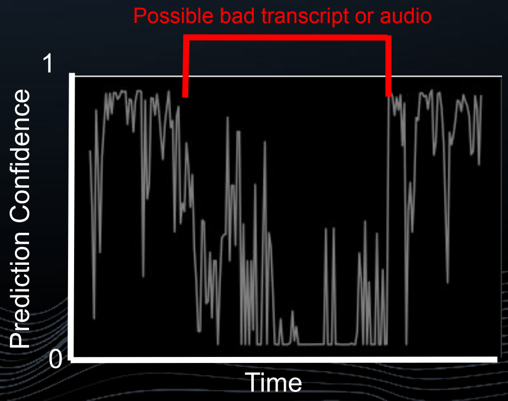
- Some signals of pathological data (e.g. incorrect transcription or corruption by noise):

Too Many Phonemes Per Unit Time

Transcript: Hey, guys. How's it going? It's Justine, and I am so excited because today I finally got my hands on the new Samsung galaxy s nine, and a huge thank you to Samsung for sending ....

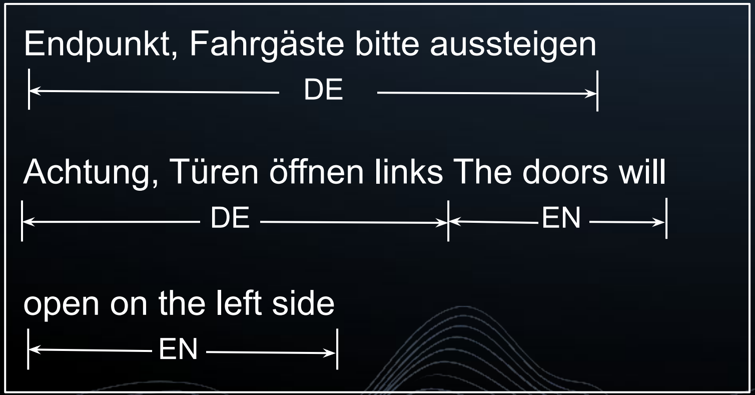


Low Confidence Model Predictions



Filtering Out-of-Domain Data

Language Detection on Transcripts



Endpunkt, Fahrgäste bitte aussteigen

DE

Achtung, Türen öffnen links The doors will

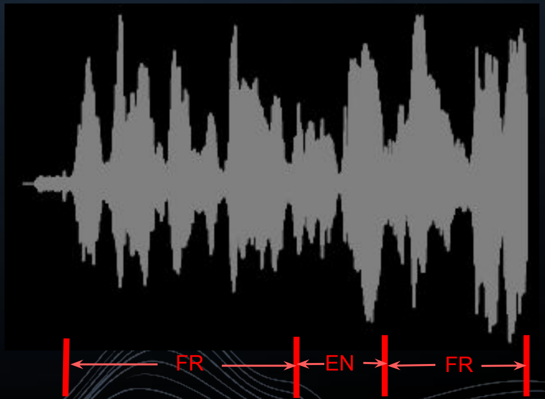
DE

EN

open on the left side

EN

Language Detection on Audio

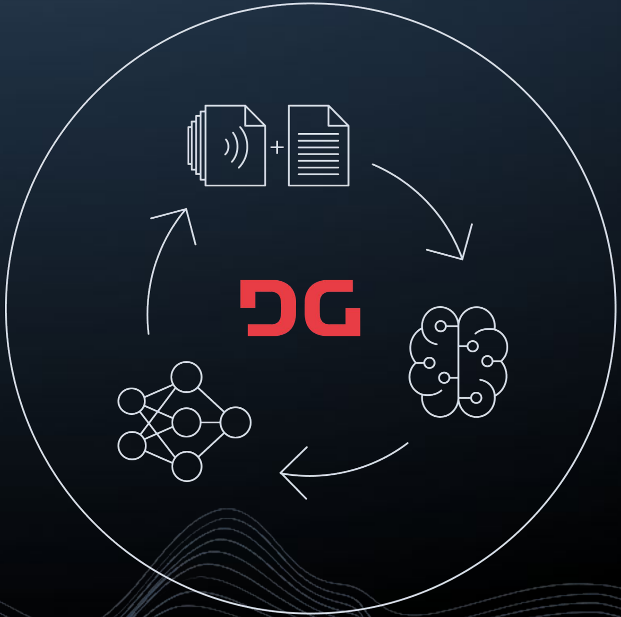


Feedback Loop with Transcription Team

Automated system to send examples where the model does poorly to be re-transcribed.

Helps transcribers refine their style guide.

If a model is performing poorly on specific words, phrases, or types of data, can self-generate targeted training data to address the problem.



Conclusion

- The ability to pull in diverse data sources, standardize the transcripts, and filter data to the task at hand is crucial.
- Data curation is an ongoing process. New data helps you to continuously improve curation tools, and the tools help you bring in new data effectively.
- Transcriptionists play an important role in targeting specific issues and specific types of data, but also need to handle more messy data that you can get at a larger scale.

DEEPGRAM

Questions

Thank you!

Matthew Lightman
matthew@deepgram.com

DEEPGRAM