DeepSpeech in 5 minutes or less

Thanks for the intro, $PERSON

SLIDE: 2021

It is 2021. Or, if you're on pandemic time, the 52nd week of March 2020.

SLIDE: 7.5 billion

There are around 7 and a half billion people on the planet right now.

SLIDE: 7139

And they speak about 7000 languages

SLIDE: 200

But only around 200 of those languages have some form of speech recognition support - that is, only around 200 language have supporting technology that allows transcription.

SLIDE: 55 of 135

And of those 200 or so languages, the majority are languages spoken in affluent or rich countries. For example, of the 135 languages supported by Google speech to text, 55 are variants of just three languages - english, spanish and arabic.

SLIDE: Indigenous languages

Indigenous languages don't get much of a look in, which is disappointing considering some Indigenous languages are spoken by millions of people - like Kinyarwanda, spoken by 12 million people in Rwanda.

This is one of the reasons why the United Nations declared 2019 the Year of Indigenous Languages, and has declared 2022-2031 the Decade of Indigenous Languages.

SLIDE: What about DeepSpeech

So what does this have to do with DeepSpeech?

Great question, I'm so glad you asked

SLIDE: Two functions

DeepSpeech has two key functions. It can be used for *inference*, which is where speech recognition is performed from a *trained model*. DeepSpeech can also be used for *training* that model.

The way DeepSpeech trains a model uses a new type of algorithm.

SLIDE: seq2seq

DeepSpeech is a *sequence to sequence* speech recognition engine. This means it doesn't use *phonemes*.

What are *phonemes*? Also a great question.

SLIDE: Phonemes

Phonemes are the building blocks of sound in a language. For example, the English language has 44 phonemes, and every word in English is made up of one or more of these *phonemes*.

SLIDE: acoustic model

Earlier types of speech recognition engines used two types of *model*. They would first use an *acoustic model* which recognised *phonemes* from a spoken phrase - like this.

SLIDE: language model

and then they would use a *language* model to turn the *phonemes* into a word, like this.

Sequence to sequence algorithms, like DeepSpeech, learn to

SLIDE: seq2seq