deeplearning.ai

# Audio data

# Speech recognition

One of the most exciting developments where sequence to sequence models has been the rise of very accurate speech recognition.
Give a sense these sequence to sequence models are applied to audio data, such as the speech.

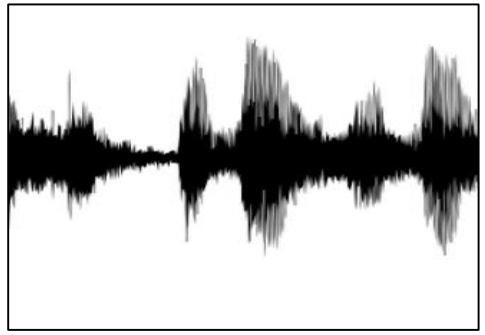
# Speech recognition problem

$x$

audio clip

$y$

transcript



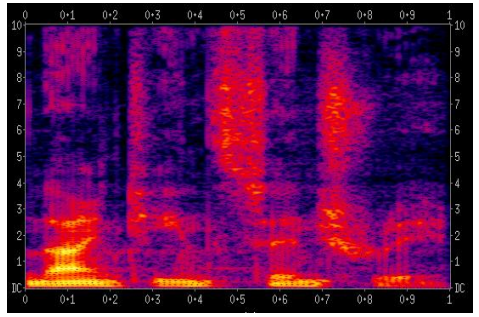
basically air pressure against time.

"the quick brown fox"

And because even the human ear doesn't process raw wave forms, but the human ear has physical structures that measures the amounts of intensity of different frequencies there is.

phonemes: de kwik braun

A common pre-processing step for audio data is to run your raw audio clip and generate a spectrogram.



So this is a plot where the horizontal axis is time, and the vertical axis is frequencies, and the intensity of different color shows the amount of energy. So, how loud is the sound at different frequencies, at different times?

And so these types of spectrograms, or you might also hear people talk about the false blank outputs, is often commonly applied pre-processing step before audio is passing into an learning algorithm.

300h

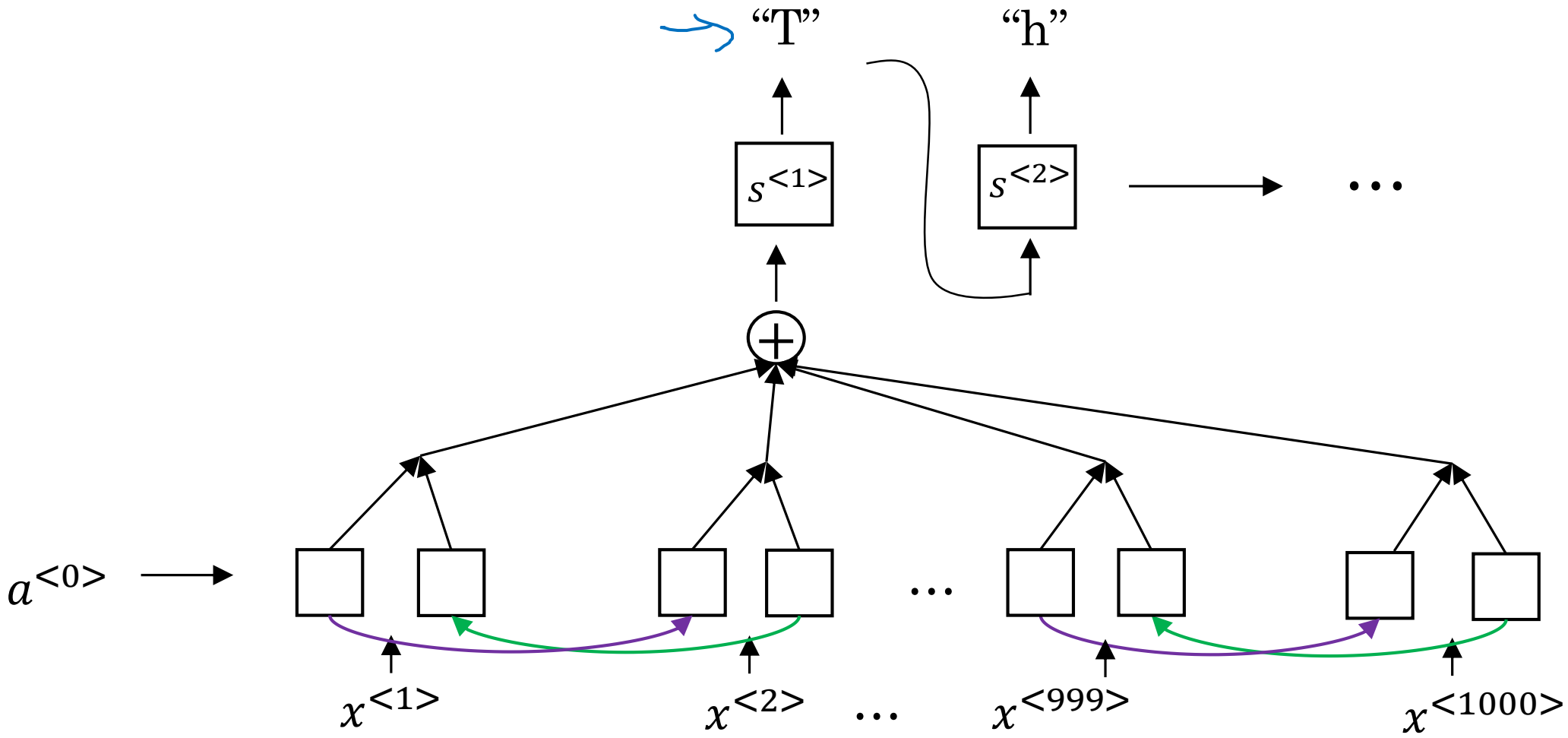
3000h

100,000h

And the human ear does a computation pretty similar to this pre-processing step.

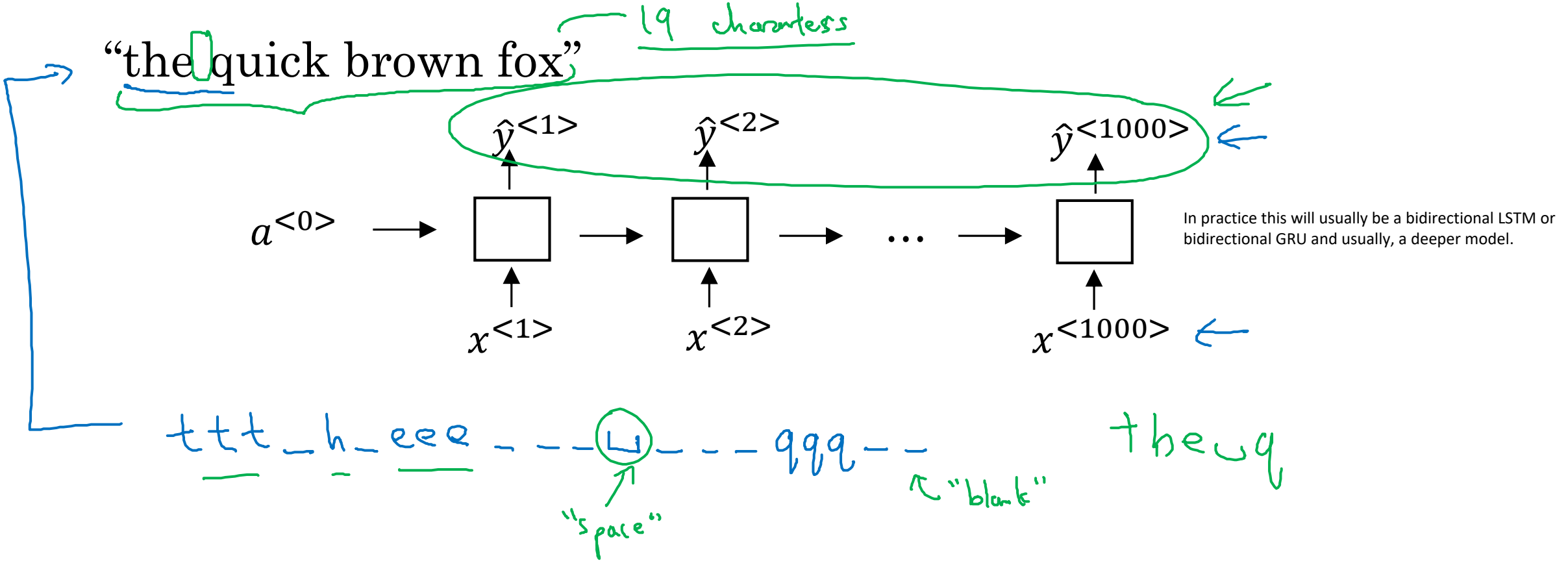
Andrew Ng

# Attention model for speech recognition



# CTC cost for speech recognition

(Connectionist temporal classification)




Basic rule: collapse repeated characters not separated by "blank"

[Graves et al., 2006. Connectionist Temporal Classification: Labeling unsegmented sequence data with recurrent neural networks]