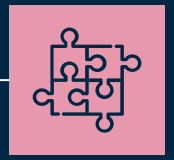
Listen, Attend, and Spell

Mahshid Alinoori

TABLE OF CONTENTS

01 Intro

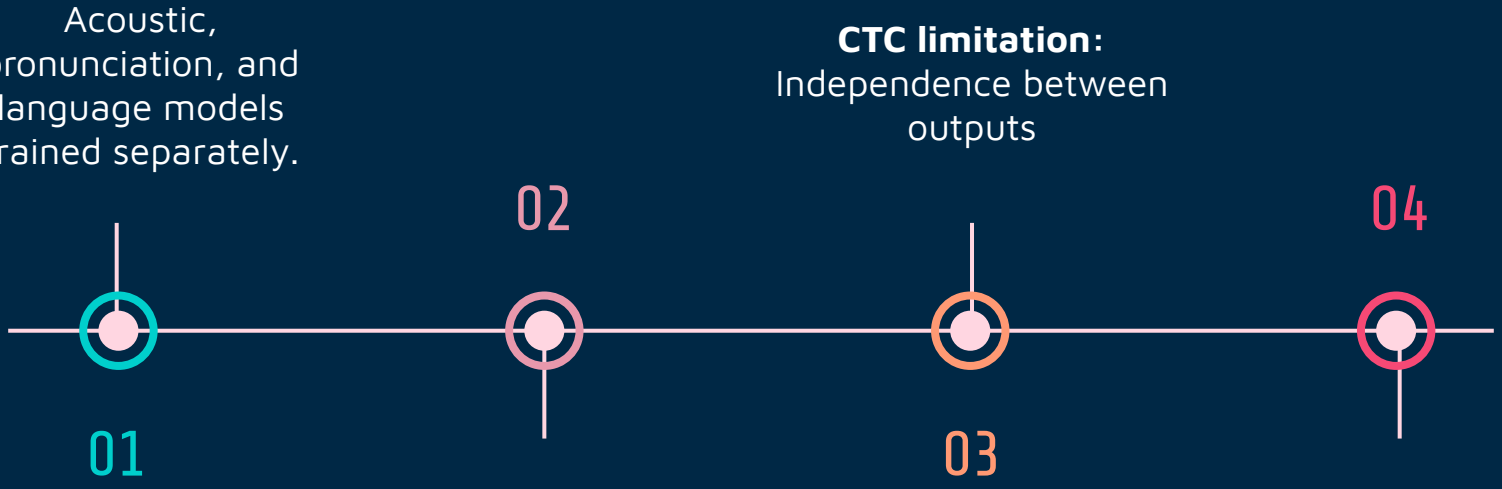


02 Model

03 Experiments

04 Conclusion

Intro


Acoustic, pronunciation, and language models trained separately.
01
Emergence of end-to-end transcription systems: CTC and seq2seq.
02
CTC limitation: Independence between outputs
03
Seq2seq limitation: Only applied to phoneme sequences
04

Intro

LAS to the rescue!

05

A seq2seq with attention: Transcribes an audio signal sequence to word sequence

06

The encoder/listener: Pyramidal Bidirectional LSTM

07

The decoder/speller: LSTM

08

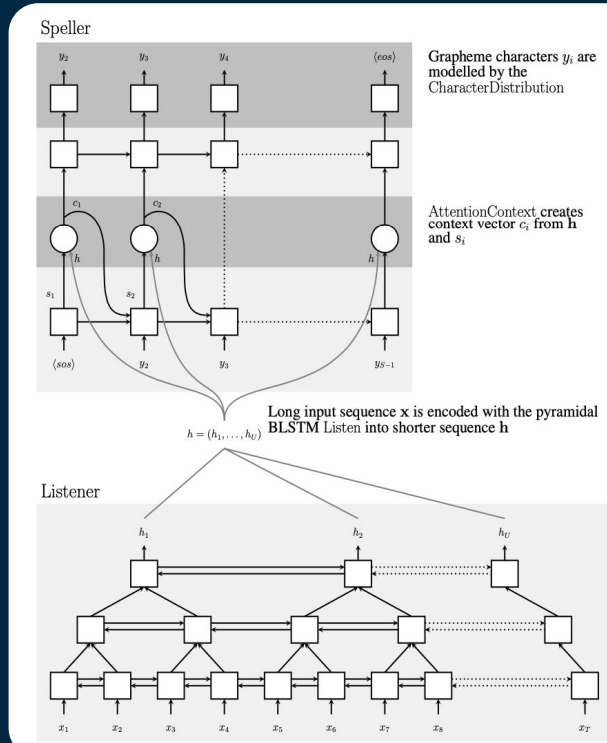
Model Architecture

Input X:

- Sequence of filter bank spectra features

Output Y:

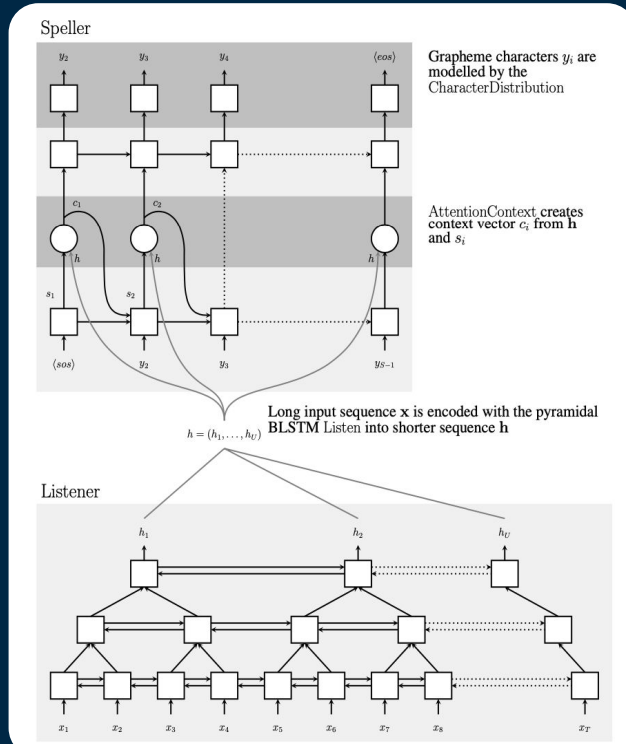
- sequence of alphanumeric characters + some punctuations
- padded by <sos> and <eos>
- modelled by $P(Y|X) = \pi P(y_i|X, y_{<i})$



Model Architecture

Listener:

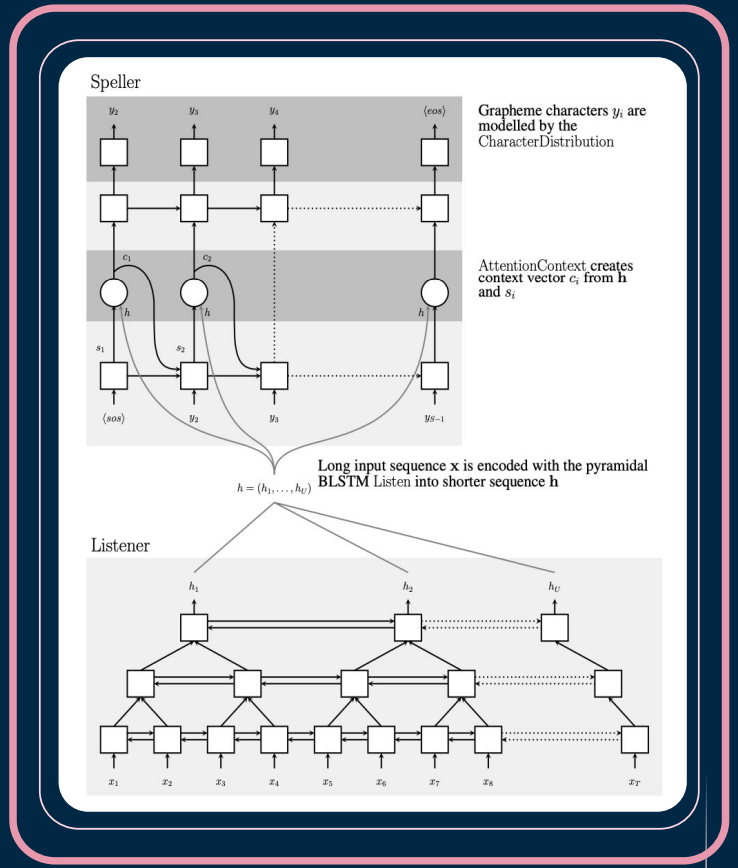
- Acoustic model encoder
- Implemented as pyramidal BLSTM
- Pyramidal structure concatenates the output of at consecutive steps and reduces the length of the encoded input and computational complexity
- 3 pBLSTMs on the top of the first layer BLSTM



Model Architecture

Attend and Spell:

- Attention-based LSTM transducer
- Outputs the distribution $P(y_i|X, y_{<i})$ using MLP and softmax on decoder state and context vector
- **Decoder state s_i:** Output of a 2 layer LSTM and a function of the previous decoder state, the previous context vector and previously generated character
- **Context vector c_i:** Output of an attention mechanism on decoder state and the listener's output h



Learning Details

Sampling Trick	Not always feeding the ground truth transcription for the next step
Decoding	Left-to-right beam search with an optional dictionary
Rescoring	Using language model to rescore the beams

Experiments and results

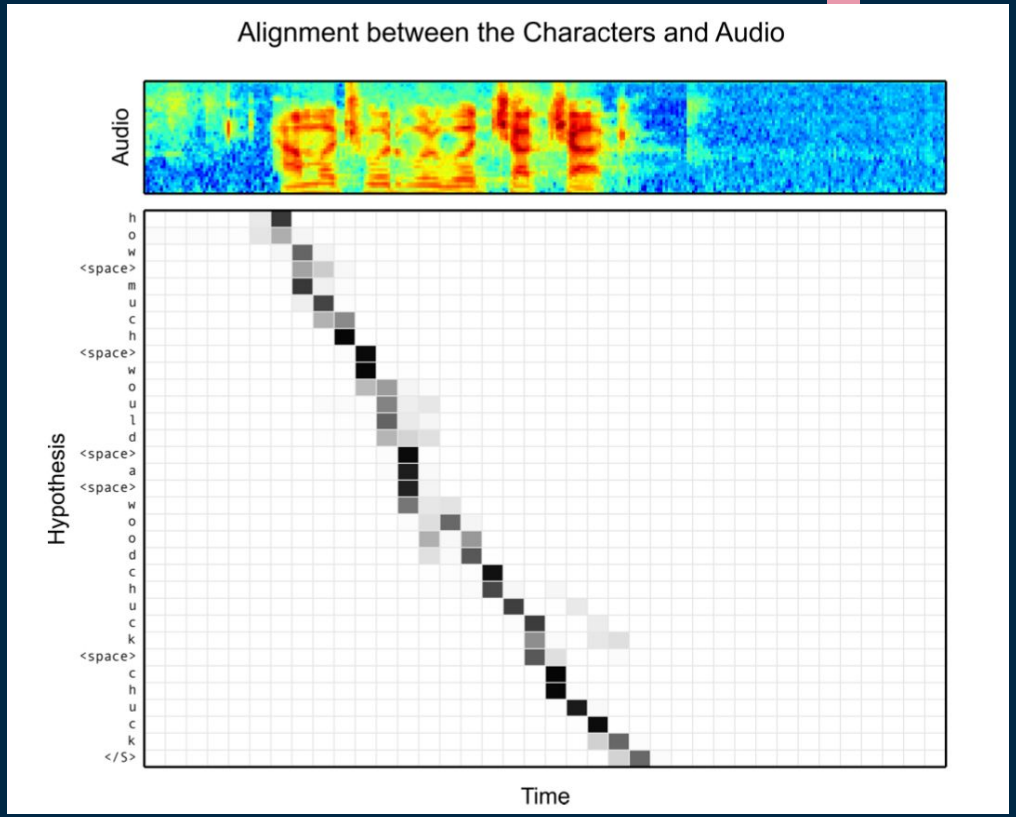
Dataset: 3 millions of Google voice search (2000 hrs)
Data Augmentation: Adding different noise and reverberation
Features: 40 dimensional Mel filter bank features

WER: A metric for word level analysis based on Levenshtein distance
Training duration: Two weeks
Tests: Clean test and noisy test

Without language model:
- WER 16.2% and 19%
- Sampling trick: WER 14.1% and 16.5%

Rescoring with language model:
- WER 12.6% and 14.7%
- Sampling trick: WER 10.3% and 12%

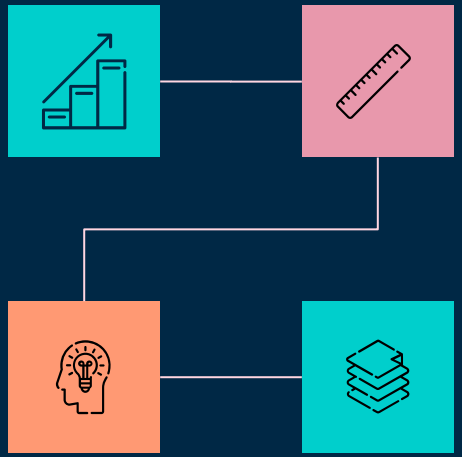
How much would a woodchuck chuck



Considerable Factors

Beam Width:

Significant WER improvement by increasing to 16



Utterance Length:

Longer utterance have higher error rate

Word Recall:

Dependant on word frequency and acoustic uniqueness

Multiple Spelling Variants:

"Call triple a roadside assistance"

aaa vs triple a

Conclusion

An end-to-end transcription system applying attention to seq2seq approach

No Attention → Overfitting
No sampling trick → Overfitting
Non-pyramidal → Slow convergence

Deep LSTMs as the listener and the speller with content-based attention

Capable of generating multiple spellings for a single input

THANKS

CREDITS: This presentation template was created by Slidesgo,
including icons by Flaticon, and infographics & images by Freepik