A Spelling Correction Model for End-to-end Speech Recognition

Jinxi Guo[1], Tara Sainath[2], **Ron Weiss**[2]

[1]*Electrical and Computer Engineering, University of California, Los Angeles, USA*

[2]*Google*

ICASSP 2019, Brighton, UK

Motivation

- End-to-end ASR models...
  - e.g. "Listen, Attend, and Spell" sequence-to-sequence model [Chan et al, ICASSP 2016]
- are trained on fewer utterances than conventional systems
  - many fewer audio-text pairs compared to text examples used to train language models
- tend to make errors on proper nouns and rare words
  - doesn't learn how to spell words which are underrepresented in the training data
- but do a good job recognizing the underlying acoustic content
  - many errors are homophonous to the ground truth

Listen, Attend, and Spell (LAS) errors

- misspells proper nouns
- replaces words with near homophones
- sometimes inconsistently

Librispeech

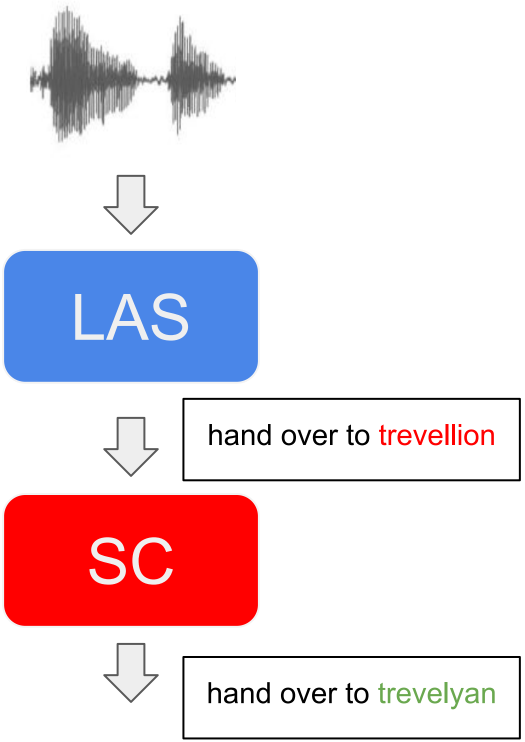
Ground Truth	LAS Output
hand over to trevelyan on trevelyan's arrival	hand over to **trevellion** on **trevelyin's** arrival
a wandering tribe of the blemmyes	a wandering tribe of the **blamies**
a wrangler's a wrangler answered big foot	a **ringleurs** a **angler** answered big foot

Can incorporate a language model (LM) trained on large text corpus [Chorowski and Jaitly, Interspeech 2017], [Kannan et al, ICASSP 2018]

Proposed Method

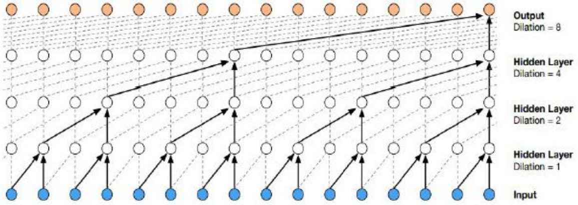
- Pass ASR hypotheses into **Spelling Correction** model
  - Correct recognition errors directly
  - or create a richer n-best list by correcting each hyp in turn
- Essentially text-to-text machine "translation" or conditional language model

- Challenge: Where to get training data?
  - *Simulate* recognition errors using large text corpus
  - Synthesize speech with TTS
  - Pass through LAS model to get hypotheses
  - Training pair: hypothesis -> Ground-truth transcript



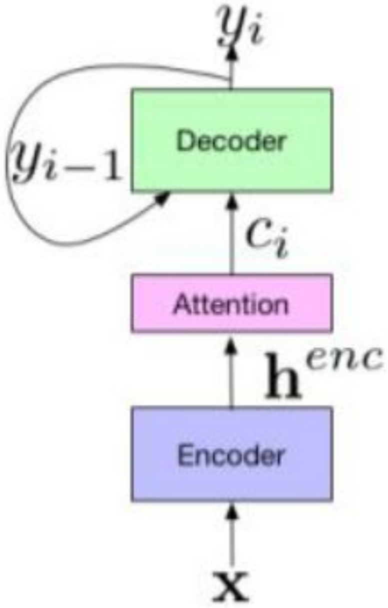
Experiments: Librispeech

- Speech
  - Read speech, long utterances
  - Training: 460 hours clean + 500 hours “other” speech
    - ~180k utterances
  - Evaluation: dev-clean, test-clean (~5.4 hours)
- Text (LM-TEXT)
  - Training: 40M sentences
- Synthetic speech (LM-TTS)
  - Synthesize speech from LM-TEXT (~60k hours) using single-voice Parallel WaveNet TTS system [Oord et al, ICML 2018]



Baseline recognizer

- Based on Listen, Attend, and Spell (LAS): attention-based encoder-decoder model
- log-mel spectrogram + delta + acceleration features
- 2x convolutional + 3x bidirectional LSTM encoder
- 4-head additive attention
- 1x LSTM decoder
- 16k wordpiece outputs



WER	DEV	TEST
LAS baseline	5.80	6.03

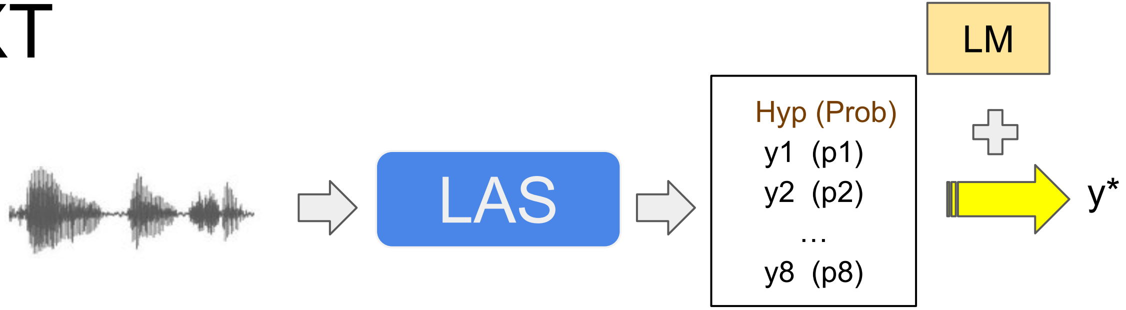
Methods for using text-only data

1. **Train LM on LM-TEXT**
   - **rescore baseline LAS output with a language model**

2. Train recognizer on LM-TTS
   - incorporate synthetic speech into recognizer training set

3. Train Spelling Corrector (SC) on decoded LM-TTS
   - train on recognition errors made on synthetic speech

Train LM on LM-TEXT



- 2 layer LSTM language model
- 16K wordpiece output vocabulary
- Rescore N-best list of 8 hyps

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \log P(\mathbf{y}|\mathbf{x}) + \lambda \log P_{LM}(\mathbf{y})$$

WER	DEV	TEST
LAS	5.80	6.03
LAS → LM (8)	**4.56** (21.4%)	**4.72** (21.7%)

LM rescoring gives significant improvement over LAS

Methods for using text-only data

1. Train LM on LM-TEXT
    - rescore baseline LAS output with a language model

2. **Train recognizer on LM-TTS**
    - **incorporate synthetic speech into recognizer training set**

3. Train Spelling Corrector (SC) on decoded LM-TTS
    - train on recognition errors made on synthetic speech

Train recognizer on LM-TTS

- Same LAS model, more training data
  - 960-hour speech + 60k-hour synthetic speech
  - "back-translation" for speech recognition [Hayashi et al, SLT 2018]
  - Each batch: 0.7*real + 0.3*LM-TTS

WER	DEV	TEST
LAS baseline	5.80	6.03
LAS-TTS	**5.68**	**5.85**
LAS → LM (8)	4.56	4.72
LAS-TTS → LM (8)	**4.45**	**4.52**

Training with combination of real and LM-TTS audio gives improvement before and after rescoring

Methods for using text-only data

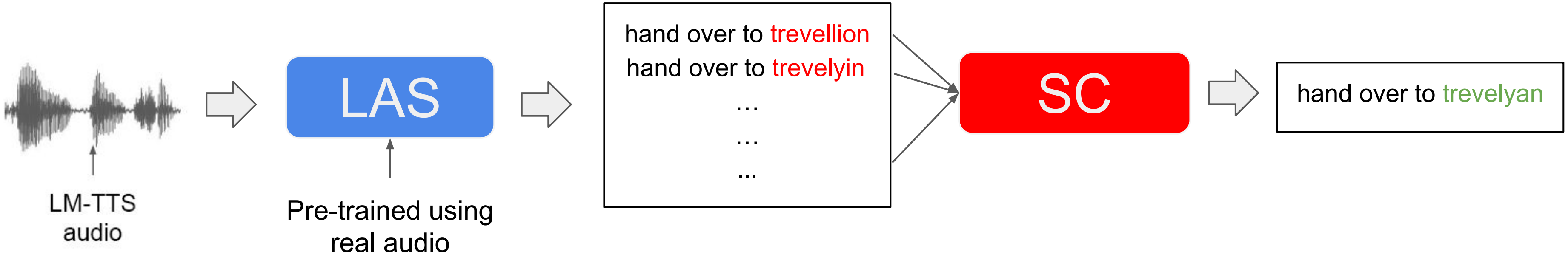
1. Train LM on LM-TEXT
    - rescore baseline LAS output with a language model

2. Train recognizer on LM-TTS
    - incorporate synthetic speech into recognizer training set

3. **Train Spelling Corrector (SC) on decoded LM-TTS**
    - **train on recognition errors made on synthetic speech**

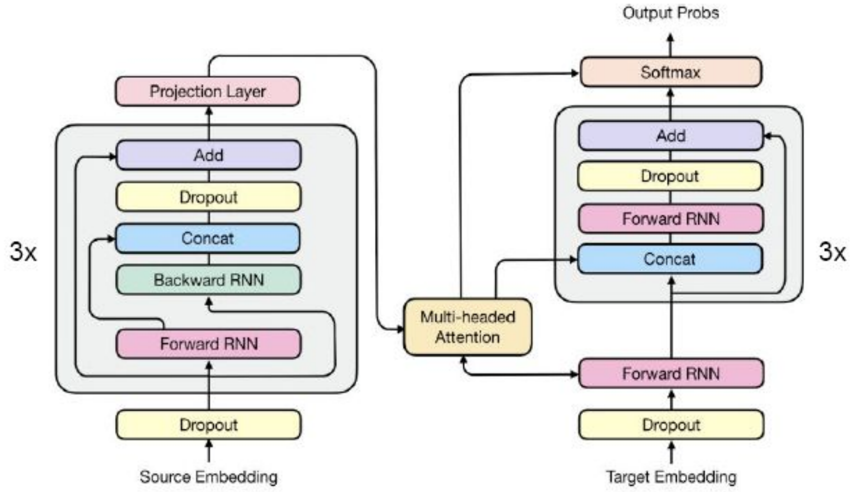
Train Spelling Corrector (SC) on decoded LM-TTS

- Training data generation
  - Baseline LAS model trained on real speech
  - Decode 40M LM-TTS utterances
    - N-best (8) list after beam-search
  - Generate text-text training pairs:
    - each candidate in the N-best list -> ground truth transcript



Model architecture

- Based on RNMT+ [Chen et al, ACL 2018]
- 16k wordpiece input/output tokens
- Encoder: 3 bidirectional LSTM layers
- Decoder: 3 unidirectional LSTM layers
- 4-head additive attention

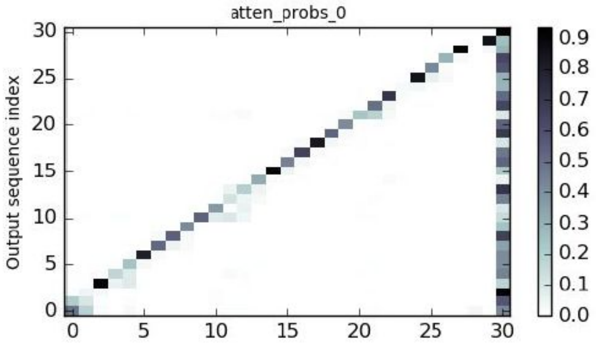


LAS → SC: Correct top hypothesis

- Directly correct the top hypothesis

WER	DEV	TEST
LAS baseline	5.80	6.03
LAS → SC (1)	**5.04** (13.1%)	**5.08** (15.8%)

- Attention weights
    - Roughly monotonic
    - Attends to adjacent context at recognition errors

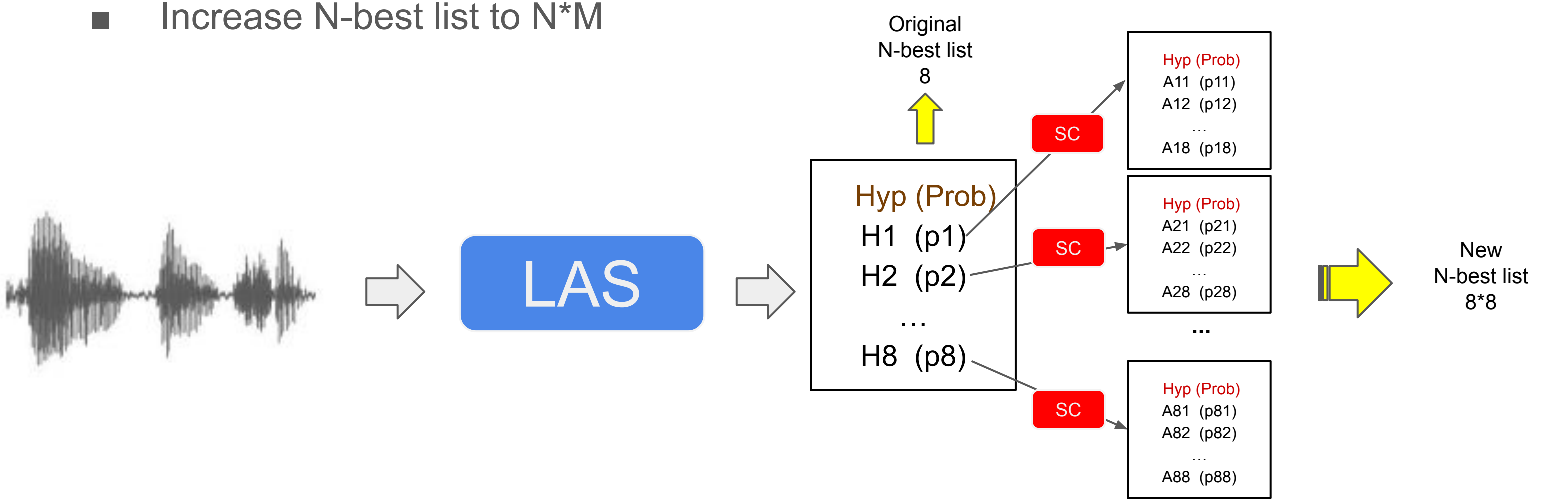


Directly applying SC to LAS top hypothesis shows clear improvement

LAS → SC: Correct N-best hypotheses

- Generate expanded N-best list
  - LAS N-best list lacks diversity
  - Pass each of N candidates to SC
    - Generate M alternatives for each one
    - Increase N-best list to N*M

ORACLE WER	DEV	TEST
LAS baseline	3.11	3.28
LAS → SC (1)	3.01	3.02
LAS → SC (8)	1.63	1.68



LAS → SC: Correct N-best hypotheses: Results

- Rescore expanded N-best list, tuning weights on dev

$$y^* = \arg\max_y \alpha\, p_{LAS}(y) + \beta\, p_{SC}(y) + \lambda\, p_{LM}(y)$$

WER	DEV	TEST	
LAS	5.80	6.03	5.26
LAS → SC (1)	5.04 (13.1%)	5.08 (15.8%)	**3.45** (34.0%)
LAS → LM (8)	4.56	4.72	3.98
LAS → SC (8) → LM (64)	**4.20** (27.6%)	**4.33** (28.2%)	**3.11** (40.9%)

Large improvement after rescoring expanded N-best list, outperforms LAS → LM

SC Train/Test mismatch

- Mismatch between recognition errors on real and TTS audio
  - Synthetic speech has clear pronunciation
    -> LAS makes fewer substitution errors

WER	DEV	TEST	**DEV-TTS**
LAS	5.80	6.03	5.26
LAS → SC (1)	5.04 (13.1%)	5.08 (15.8%)	**3.45** (34.0%)
LAS → LM (8)	4.56	4.72	3.98
LAS → SC (8) → LM (64)	**4.20** (27.6%)	**4.33** (28.2%)	**3.11** (40.9%)

Results on DEV-TTS show potential of SC when errors are matched between train and test

Multistyle Training (MTR)

- Increase SC training data variability
- Add noise and reverberation to LM-TTS [Kim et al, Interspeech 2017]
- Train on LM-TTS clean + MTR
  - total of 640M training pairs

WER	DEV	TEST
LAS baseline	5.80	6.03
LAS → SC (1)	5.04 (13.1)	5.08 (15.8%)
LAS → SC-MTR (1)	**4.87** (16.0%)	**4.91** (18.6%)
LAS → LM (8)	4.56	4.72
LAS → SC (8) → LM (64)	4.20 (27.6%)	4.33 (28.2%)
LAS → SC-MTR (8)→ LM (64)	**4.12** (29.0%)	**4.28** (29.0%)

MTR makes TTS audio more realistic and generates noisier N-best list with better matched errors

Example corrections

- Corrects proper nouns, rare words, tense errors

Reference	LAS baseline	LAS → LM (8)	LAS → SC (8) → LM (64)
ready to hand over to trevelyan on trevelyan's arrival in england	ready to hand over to trevellion on trevelyin's arrival in england	ready to hand over to trevellion on trevelyan's arrival in england	ready to hand over to trevelyan on trevelyan's arrival in england
has countenanced the belief the hope the wish that the ebionites or at least the nazarenes	has countenance the belief the hope the wish that the epeanites or at least the nazarines	has countenance the belief the hope the wish that the epeanites or at least the nazarines	has countenanced the belief the hope the wish that the ebionites or at least the nazarenes
a wandering tribe of the blemmyes or nubians	a wandering tribe of the blamies or nubians	a wandering tribe of the blamis or nubians	a wandering tribe of the blemmyes or nubians

Example incorrections

- Spelling corrector sometimes introduces errors

Reference	LAS baseline	LAS → LM (8)	LAS → SC (8) → LM (64)
a laudable regard for the <u>honor</u> of the first proselyte	a laudable regard for the <u>honor</u> of the first proselyte	a laudable regard for the <u>honor</u> of the first proselyte	a laudable regard for the <u>honour</u> of the first proselyte
ambrosch he <u>make</u> good farmer	ambrosch he <u>may</u> good farmer	ambrose he <u>make</u> good farmer	ambrose he <u>made</u> good farmer

Summary

- **Spelling correction** model to correct recognition errors
- Outperforms LM rescoring alone by expanding N-best list
- MTR data augmentation improves SC model
  - Overall ~29% relative improvement
- Future work: better strategies for creating better matched SC training data

WER	DEV	TEST
LAS baseline	5.80	6.03
LAS-TTS	5.68	5.85
LAS → SC (1)	5.04	5.08
LAS → SC-MTR (1)	**4.87**	**4.91**
LAS → LM (8)	4.56	4.72
LAS-TTS → LM (8)	4.45	4.52
LAS → SC (8) → LM (64)	4.20	4.33
LAS → SC-MTR (8) → LM (64)	**4.12**	**4.28**

Q&A

Thanks for your attention!

Acknowledgements:

- Zelin Wu, Anjuli Kannan, Dan Liebling, Rohit Prabhavalkar, Kazuki Irie, Golan Pundak, Melvin Johnson, Mia Chen, Zhouhan Lin, Antonios Anastasopoulos and Uri Alon

Contact: Jinxi Guo `lennyguo@gmail.com`