Multi-Dialect Speech Recognition With A Single Seq2Seq Model

Bo Li, Tara N. Sainath, Khe Chai Sim, Michiel Bacchiani,
Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yonghui Wu, Kanishka Rao

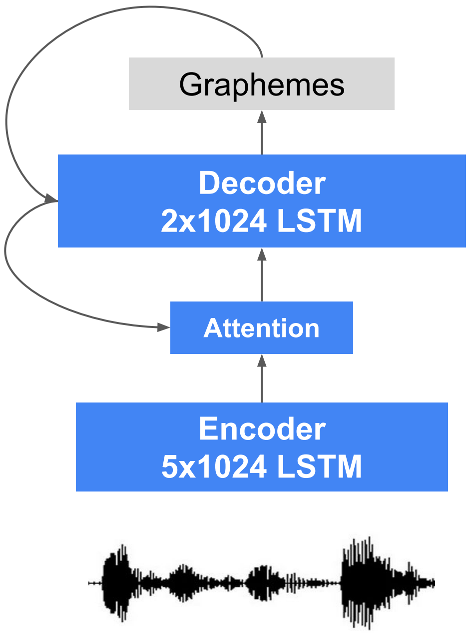
ICASSP 2018
Tue-SP-L1.1

Agenda

- Introduction
- Multi-Dialect LAS Model
  - Dialect as Output Targets
  - Dialect as Input Features
  - Dialect as Cluster Coefficients
- Experimental Evaluations
- Conclusions

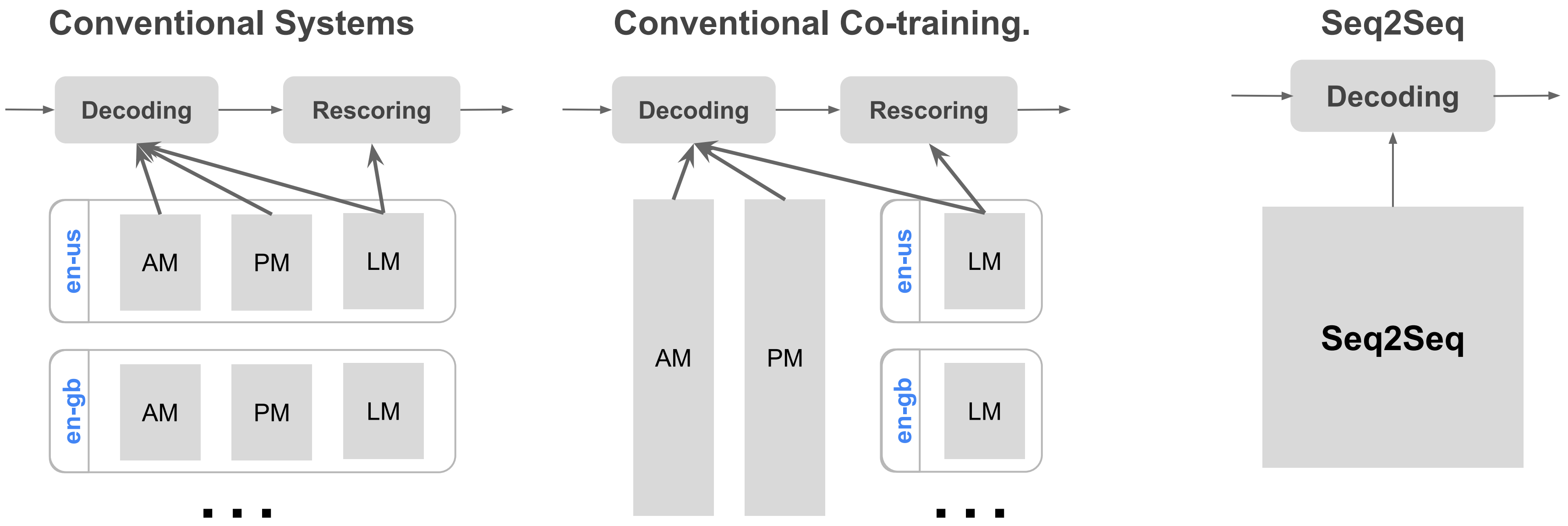
Listen, Attend & Spell - LAS [1]

- Attention-based sequence-to-sequence model
- **Jointly** learns "**acoustic**" and "**language**" model components
- Attention mechanism summarizes relevant encoder features to predict next label
- Previous label prediction is fed back into the decoder to predict the current one



[1] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015

Multi-Dialect ASR



In conventional systems, languages/dialects,
are handled with **individual AMs, PMs and LMs**.
Upscaling is becoming challenging.

A single model for all.

Google

Multi-Dialect LAS

- Modeling Simplicity
- Data Sharing
  - among dialects and model components
- Joint Optimization
- Infrastructure Simplification
  - a single model for all

Table: Resources required for building each system.

Conventional	Seq2Seq
data phoneme lexicon text normalization LM × N	data

Multi-Dialect LAS

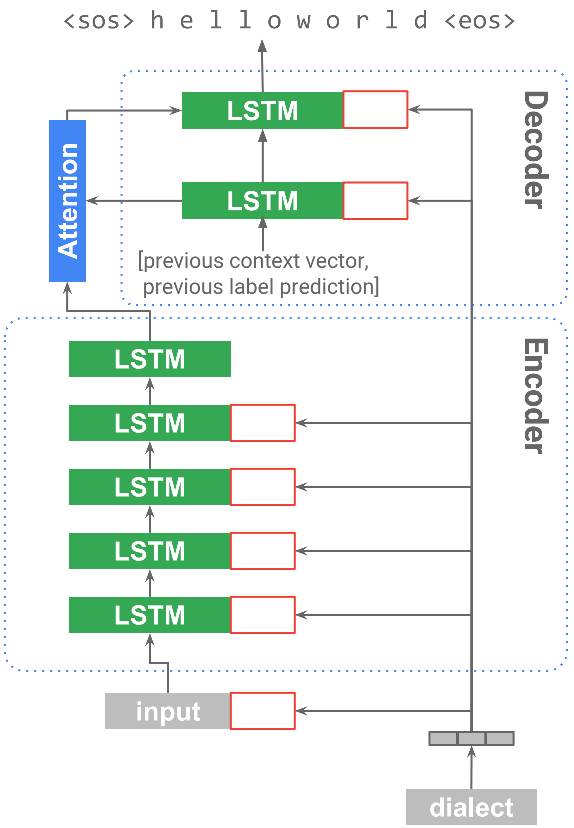
Dialect as Output Targets

- Multi-Task Learning: Joint Language ID (LID) and ASR
  - LID first, then ASR
    - "<sos> **<en-gb>** h e l l o ⊔ w o r l d <eos>"
    - LID errors may affect ASR performance
  - ASR first, then LID
    - "<sos> h e l l o ⊔ w o r l d **<en-gb>** <eos>"
    - ASR prediction is not dependent on LID prediction, not suffering from LID errors

Dialect as Input Features

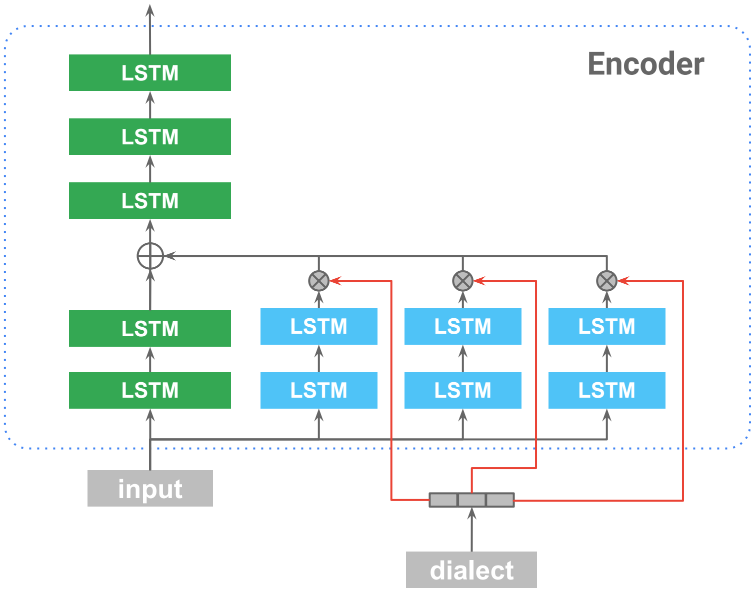
- Passing the dialect information as additional features

components		variations
encoders	→	acoustic
decoders	→	lexicon and language



Dialect Information as Cluster Coefficients

- Cluster Adaptive Training (CAT) [1] coefficients
  - more flexible model architectures
  - larger capacity in variation modeling
  - but increased model parameters



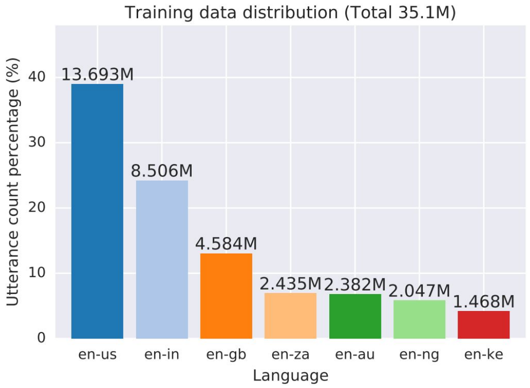
Google

[1] Tian Tan, Yanmin Qian and Kai Yu, "Cluster adaptive training for deep neural network based acoustic model", IEEE/ACM TASLP, 2016

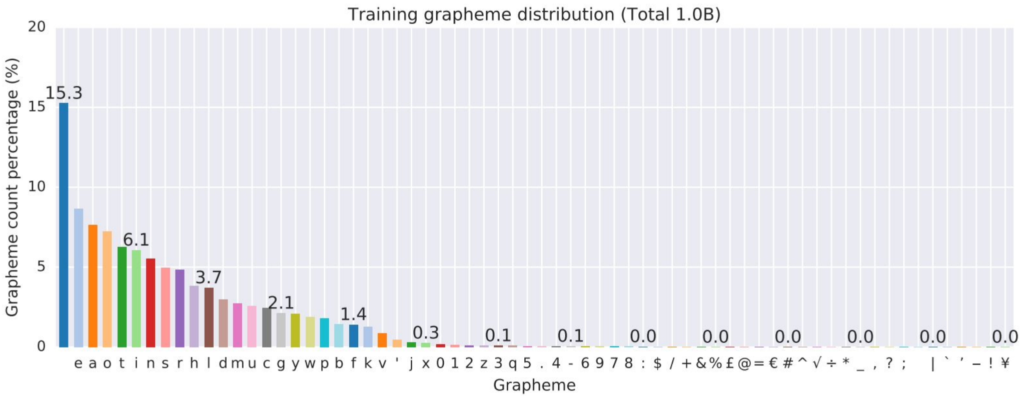
Experimental Evaluations

Task

- **7 English dialects:** `US (America), IN (India), GB (Britain), ZA (South Africa), AU (Australia), NG (Nigeria & Ghana), KE (Kenya)`



★ **unbalanced** dialect data



★ **unbalanced** target classes

Google

LAS Co-training Baselines

Dialect	US	IN	GB	ZA	AU	NG	KE
dialect-ind.	10.6	18.3	12.9	12.7	12.8	33.4	19.2
dialect-dep.	**9.7**	**16.2**	**12.7**	**11.0**	**12.1**	33.4	**19.0**

★ dialect specific **fine-tuning still wins**

★ simply pooling the data is **missing** certain dialect specific variations

LAS With Dialect as Output Targets

Dialect	US	IN	GB	ZA	AU	NG	KE
Baseline (dialect-dep.)	9.7	**16.2**	12.7	11.0	12.1	33.4	19.0
LID first	9.9	16.6	12.3	11.6	12.2	33.6	18.7
ASR first	**9.4**	16.5	**11.6**	**11.0**	**11.9**	**32.0**	**17.9**

- ★ LID error **affects** ASR
- ★ **ASR first** is better

Example target sequence

LID first	`<sos> <en-gb> h e l l o ⊔ w o r l d <eos>`
ASR first	`<sos> h e l l o ⊔ w o r l d <en-gb> <eos>`

LAS With Dialect as Input Features

Dialect		US	IN	GB	ZA	AU	NG	KE
Baseline (dialect-dep.)		9.7	16.2	12.7	11.0	12.1	33.4	19.0
encoder	**1-hot**	9.6	16.4	11.8	10.6	10.7	31.6	18.1
	emb.	9.6	16.7	12.0	10.6	10.8	32.5	18.5
decoder	**1-hot**	9.4	16.2	11.3	10.8	10.9	32.8	18.0
	emb.	9.4	16.2	**11.2**	10.6	11.1	32.9	18.0
both	**1-hot**	**9.1**	**15.7**	11.5	**10.0**	**10.1**	**31.3**	**17.4**

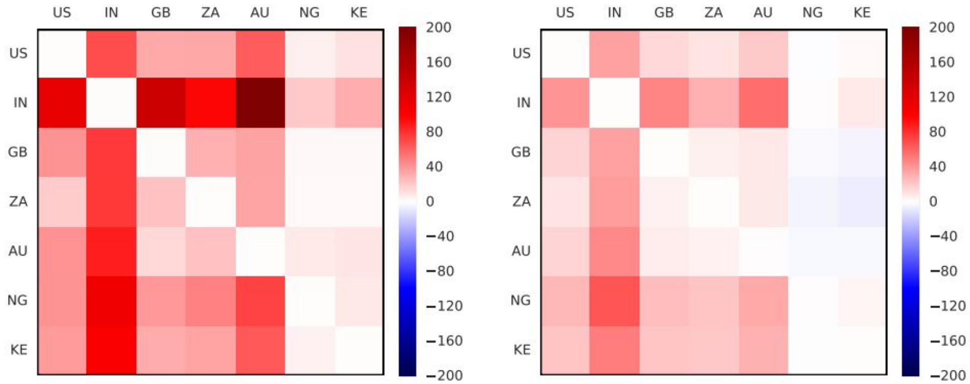
★ dialect 1-hot and embedding (emb.) performs **similarly**

★ feeding dialect to **both encoder and decoder** gives the largest gains

Google

LAS With Dialect as Input Features

Figure: Feeding different dialect vectors (rows) to the LAS encoder and decoder on different test sets (columns).



(a) Encoder (b) Decoder

★ **encoder** is more sensitive to wrong dialects → large acoustic variations

★ for **low-resource** dialects (NG, KE), the model **learns to ignore** the dialect information

LAS With Dialect as Input Features

- The dialect vector does **both AM and LM adaptation**

*Table: The number of **color**/**colour** occurrences in hypotheses on the **en-gb** test data.*

dialect vector	encoder	decoder	color (US)	colour (GB)
×	×	×	1	22
<en-gb>: [0, 1, 0, 0, 0, 0, 0]	✓	×	19	4
<en-gb>: [0, 1, 0, 0, 0, 0, 0]	×	✓	0	25
<en-us>: [1, 0, 0, 0, 0, 0, 0]	×	✓	24	0

★ dialect vector helps **encoder** to **normalize accent variations**

★ dialect vector helps **decoder** to **learn dialect-specific lexicons**

LAS With Dialect as CAT coefficients

Dialect		US	IN	GB	ZA	AU	NG	KE
Baseline (dialect-dep.)		9.7	16.2	12.7	11.0	12.1	33.4	19.0
input features (encoder)	**1-hot**	9.6	16.4	11.8	10.6	10.7	**31.6**	**18.1**
CAT coeff.	**1-hot**	9.9	17.0	12.1	11.0	11.6	32.5	18.3
	emb.	**9.4**	**16.1**	**11.7**	**10.6**	**10.6**	32.9	**18.1**

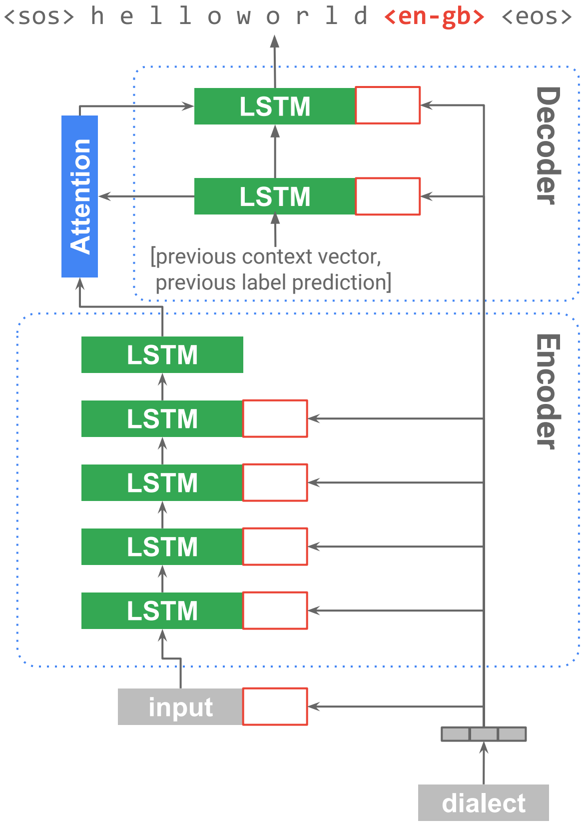
★ dialect as CAT coefficients is much better than as inputs

★ but with large model params increase (160K vs. 3M)

Final Multi-Dialect LAS

Final Multi-Dialect LAS

- output targets:
  - multi-task with ASR first
- input features:
  - feeding dialect to both encoder and decoder



Google

Final Multi-Dialect LAS

Dialect	US	IN	GB	ZA	AU	NG	KE
Baseline (dialect-dep.)	9.7	16.2	12.7	11.0	12.1	33.4	19.0
output targets (ASR first)	9.4	16.5	11.6	11.0	11.9	32.0	17.9
input features (both)	**9.1**	**15.7**	11.5	10.0	**10.1**	**31.3**	**17.4**
final	**9.1**	16.0	**11.4**	**9.9**	10.3	31.4	17.5

★ **small gains** when combining input and output

★ the final system **outperforms** the dialect-dependent models by 3.1~16.5% relatively

Conclusions

- We investigated building Multi-Dialect LAS Models with additional dialect information:
  - as additional Output Targets (multi-task learning)
  - as extra Input Vectors
  - as Cluster Adaptation Training coefficients
- We justified:
  - the feasibility of building a single LAS model to capture dialect variations
  - dialect information boosts the single model to outperform dialect dependent models.