Towards Language-Universal End-to-End Speech Recognition

Suyoun Kim[1], and Michael L. Seltzer[2]

[1]Carnegie Mellon University

[2]Facebook (formerly Microsoft AI & Research)

Microsoft

Carnegie Mellon University

ICASSP April 18, 2018

Presenter: Suyoun Kim

Outline

- Motivation of Language-universal end-to-end speech recognition
- Proposed model: language-specific gated network
- Experimental evaluation
- Conclusions

Challenges of growing language coverage of ASR systems

- There are over 6,000 languages globally

1) Conventional ASR requires **each model be trained independently**
   - Effort to train, deploy, and maintain so many models in production increases

2) For second and third tier languages, additional challenges arise
   - **Lack of sufficient training data**
   - **Lack of linguistic expertise, lexicons**

Prior work: multi-lingual acoustic models

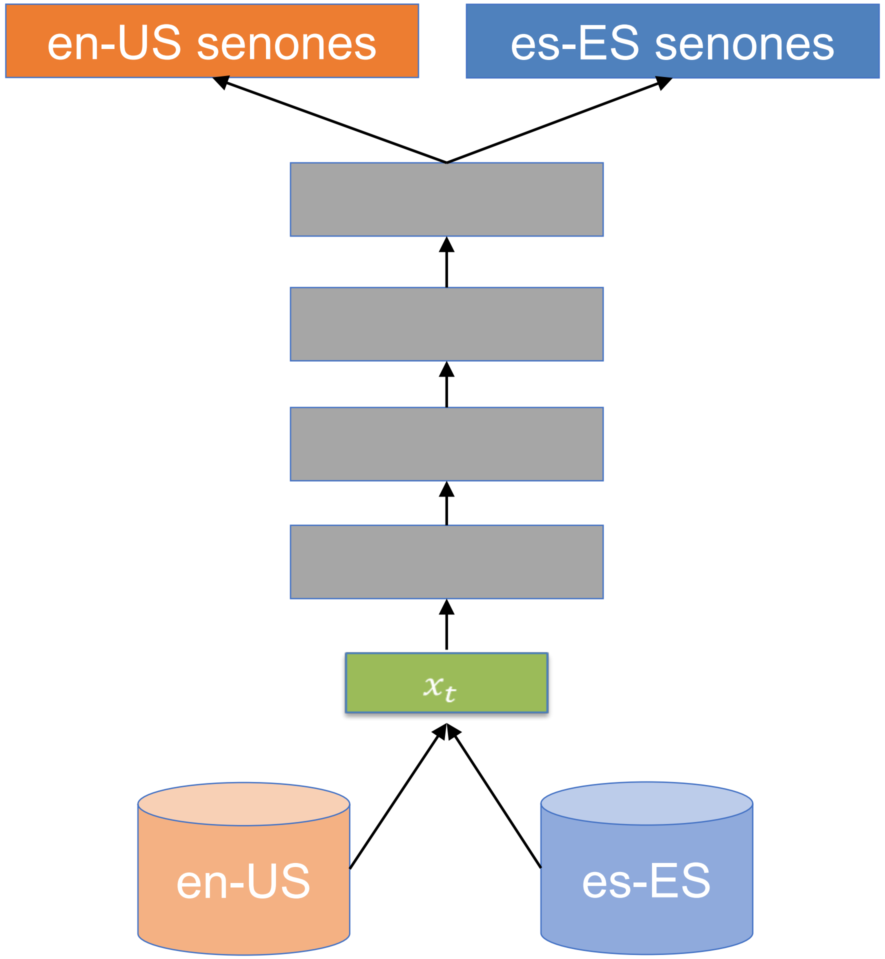
- Transfer learning approach:
  - Share language-independent lower layer(s)
  - Separate language-specific output layer(s)

✓ **Pools data to train common parameters**

✓ **Improved performance with (very) little training data**

× **Requires pronunciation lexicon**

× **Improvement diminishes with increased data**



Our model: A language-universal end-to-end ASR

- Key insights

1) End-to-end with CTC

2) Universal character set

3) Language-specific gating

Our model: A language-universal end-to-end ASR

- Key insights

1) End-to-end with CTC[1]

- No pronunciation lexicon required

- Convert a sequence of features to a sequence of graphemes rather than senones

[1] *Graves et al. 2016*

Our model: A language-universal end-to-end ASR

- Key insights

1) End-to-end with CTC

- No pronunciation lexicon required

2) Universal character set

- Single system
- Easy to maintain

3) Language-specific gating

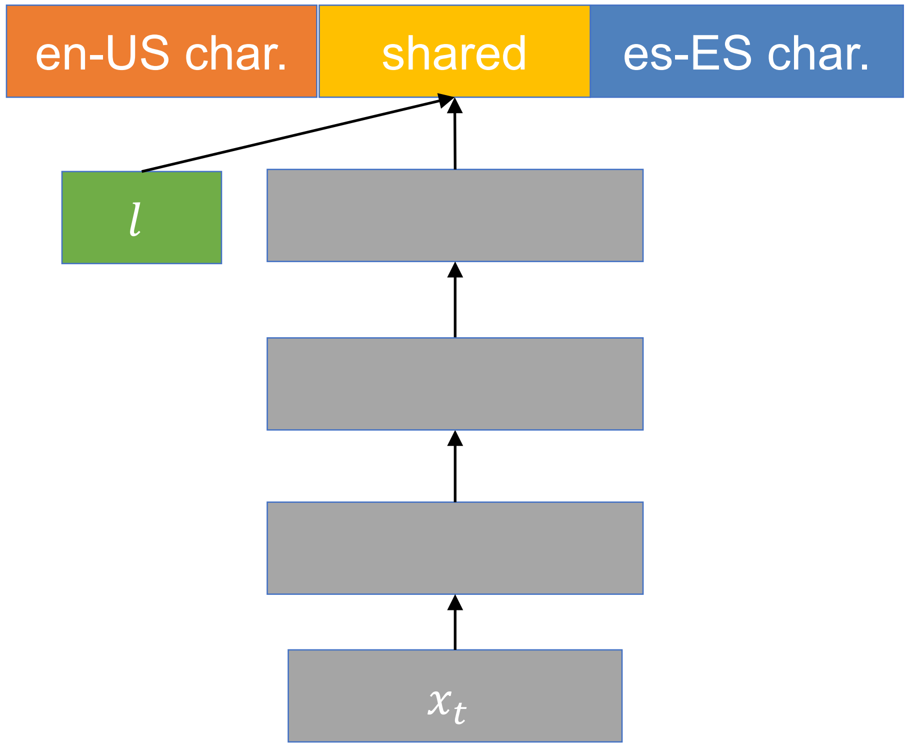
2) Use a universal character set

- Share model parameters and even output layer among languages
  - **Single system** capable of recognizing any language it has been trained on

- Assume language identity is known in training and decoding

- Mask out the activation from unwanted characters

Union Labels



“Universal keyboard” shares common characters

Experiment setup

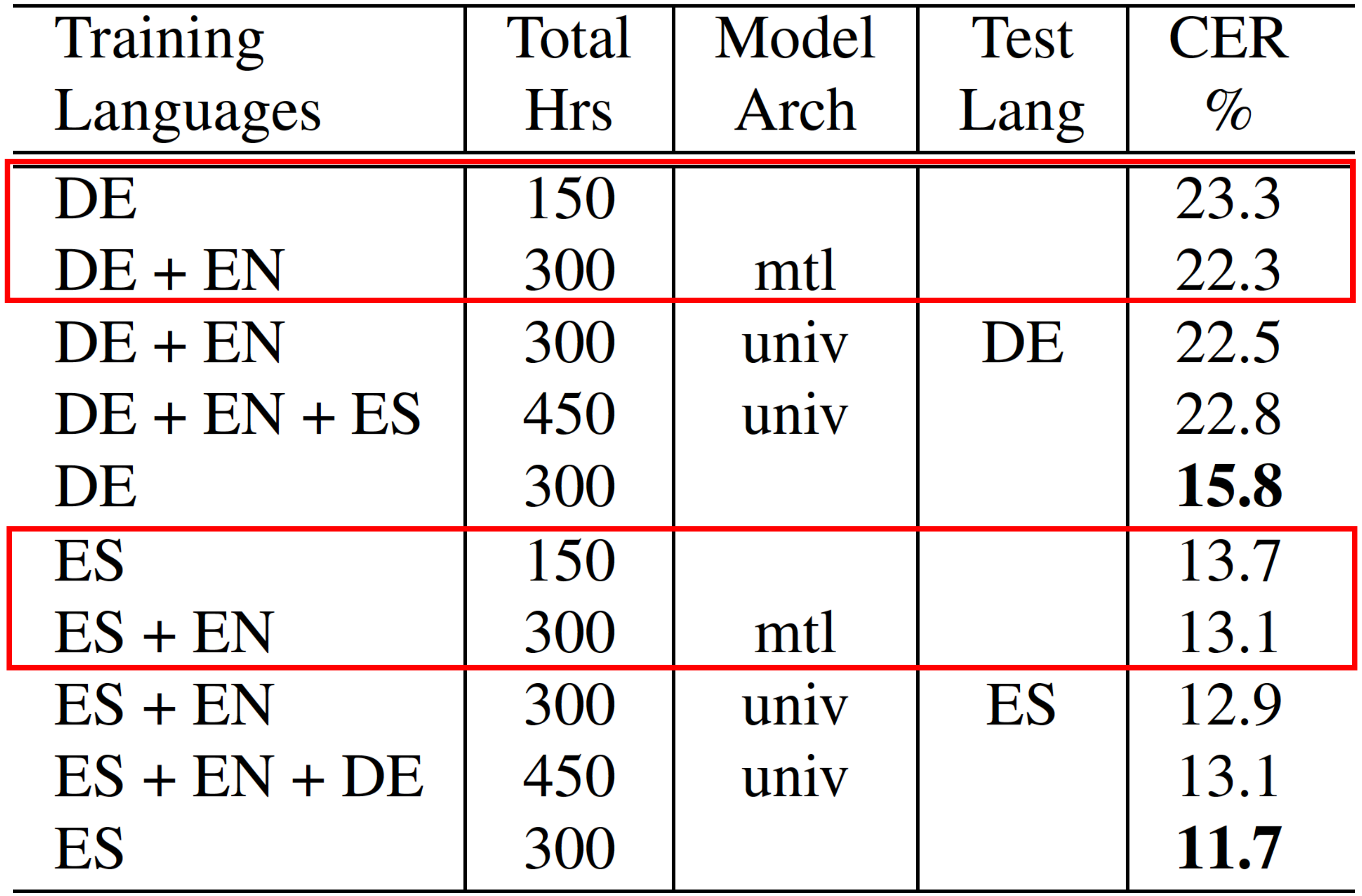
- Data
  - Cortana data in English (EN), Spanish (ES), and German (DE)
  - 150 hour training set, 10 hour dev set, 10 hour test set, per language

- Model:
  - Input: 80-dimensional log mel filterbank x 3
  - Output: characters (graphemes)[1] - EN: 81d, DE: 93d, ES: 97d
  - 4 layer BLSTM (320 cells)

- Training and Decoding
  - CTC with SGD with fixed learning rate, early stopping, random initialization
  - Greedy decoding with no explicit language model

[1] *Zweig et al., advances in all-neural speech recognition, 2016*

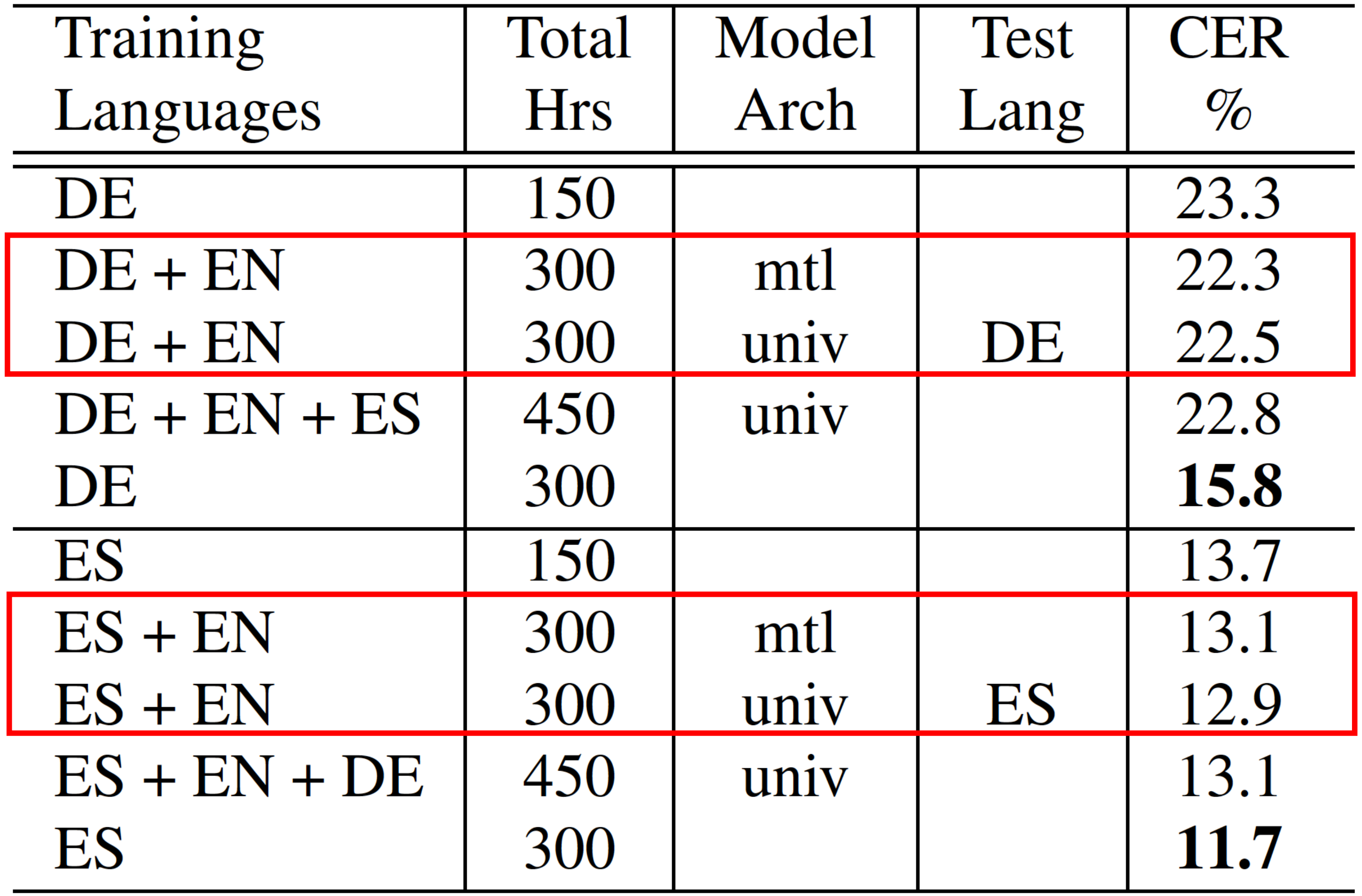
Initial evaluation:



Training Languages	Total Hrs	Model Arch	Test Lang	CER %
DE	150			23.3
DE + EN	300	mtl		22.3
DE + EN	300	univ	DE	22.5
DE + EN + ES	450	univ		22.8
DE	300			**15.8**
ES	150			13.7
ES + EN	300	mtl		13.1
ES + EN	300	univ	ES	12.9
ES + EN + DE	450	univ		13.1
ES	300			**11.7**

1. Small gain by adding different EN training source

Initial evaluation: multi-task vs. union architectures

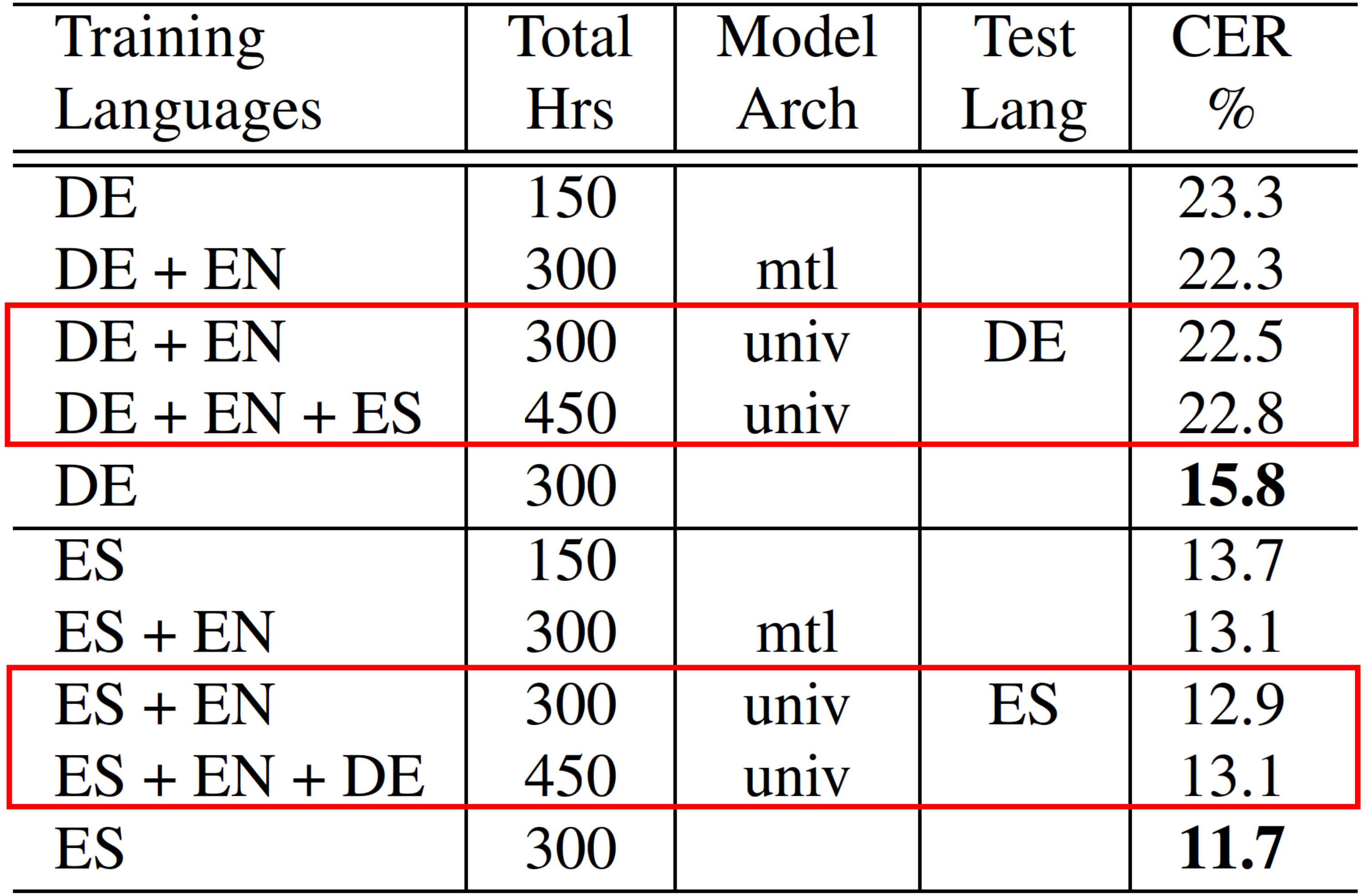


Training Languages	Total Hrs	Model Arch	Test Lang	CER %
DE	150		DE	23.3
DE + EN	300	mtl		22.3
DE + EN	300	univ		22.5
DE + EN + ES	450	univ		22.8
DE	300			**15.8**
ES	150		ES	13.7
ES + EN	300	mtl		13.1
ES + EN	300	univ		12.9
ES + EN + DE	450	univ		13.1
ES	300			**11.7**

1. Small gain by adding different EN training source

2. Separate labels (mtl) and universal labels (univ) perform comparably

Initial evaluation:
No improvement increasing from 2 langs. to 3 langs.



Training Languages	Total Hrs	Model Arch	Test Lang	CER %
DE	150		DE	23.3
DE + EN	300	mtl		22.3
DE + EN	300	univ		22.5
DE + EN + ES	450	univ		22.8
DE	300			**15.8**
ES	150		ES	13.7
ES + EN	300	mtl		13.1
ES + EN	300	univ		12.9
ES + EN + DE	450	univ		13.1
ES	300			**11.7**

1. Small gain by adding different EN training source
2. Separate labels (mtl) and universal labels (univ) perform comparably
3. **No improvement increasing from 2 languages to 3 languages**

Our model: A language-universal end-to-end ASR

- Key insights

1) End-to-end with CTC

- No pronunciation lexicon required

2) Universal character set

- Single system
- Easy to maintain

3) Language-specific gating

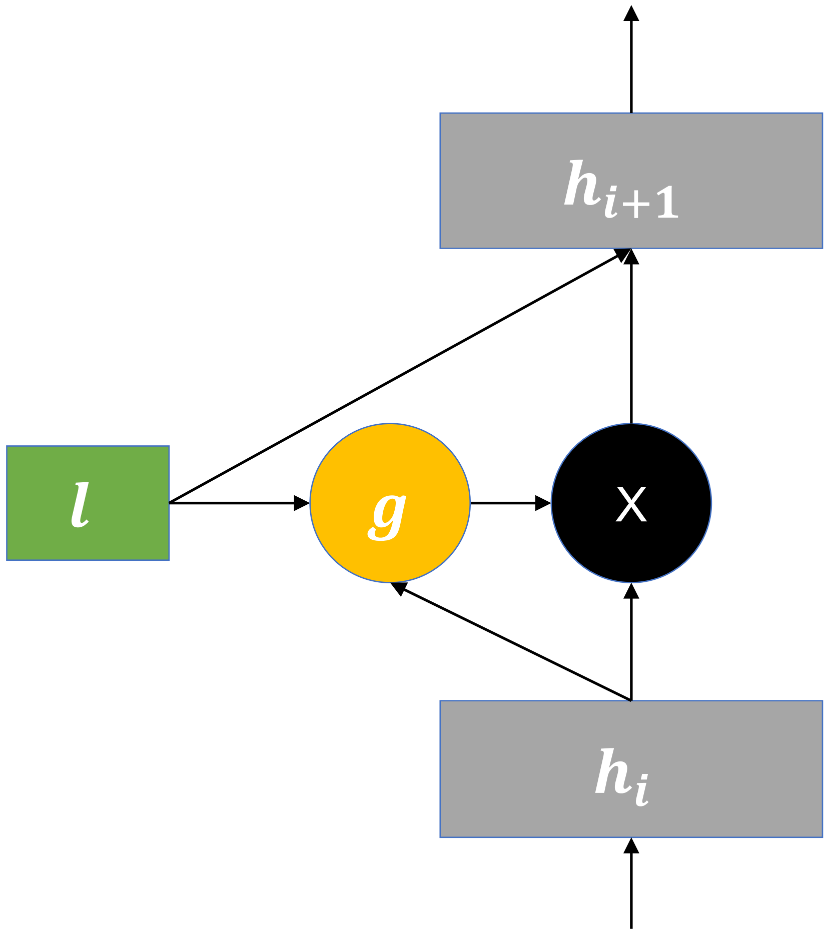
- Further improvement with more data

3) language-specific gating

- Motivation: model needs to adequately capture language-specific information
    - Adding language ID indicator gives minimal improvement

=> Add language-specific gating mechanism to each layer

- Modulate internal representations in a language-specific way
- **Fewer parameter** than ***cluster adaptive training (CAT)***[1][2]



[1] *Li et al., multi-dialect speech recognition with a single sequence-to-sequence model, 2018*
[2] *Tan et al., cluster adaptive training for deep learning network based acoustic model, 2016*

3) language-specific gating: implementation details

1. Define one-hot language indicator vector d_l

$$d_l = [0\ 0\ 1]$$

2. Compute gate for i^{th} hidden layer

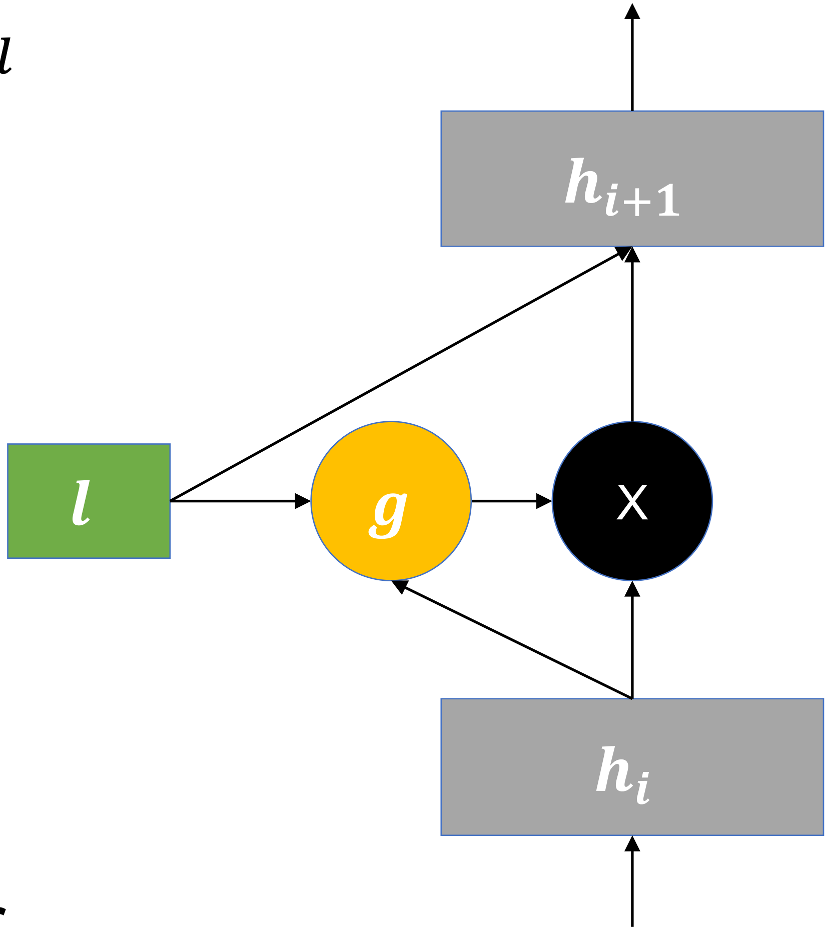
$$g(h_i, l) = \sigma(\boldsymbol{U} h_i + \boldsymbol{V} d_l + \boldsymbol{b})$$

3. Compute language-gated activation

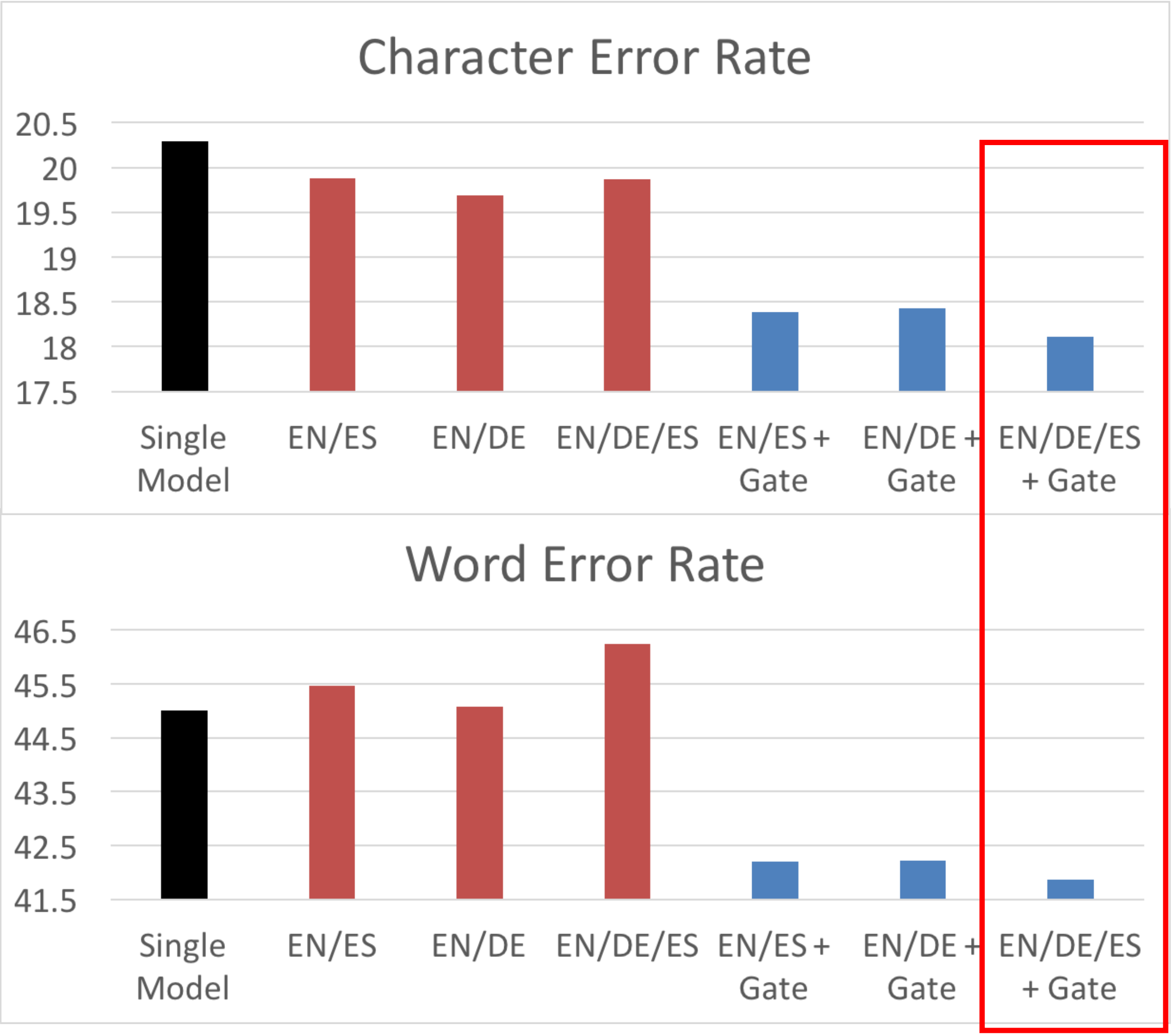
$$\hat{h}_i = g(h_i, l) \odot h_i$$

4. Gated activations and d_l input to next layer

$$\tilde{h}_i = [\hat{h}_i \colon d_l]$$

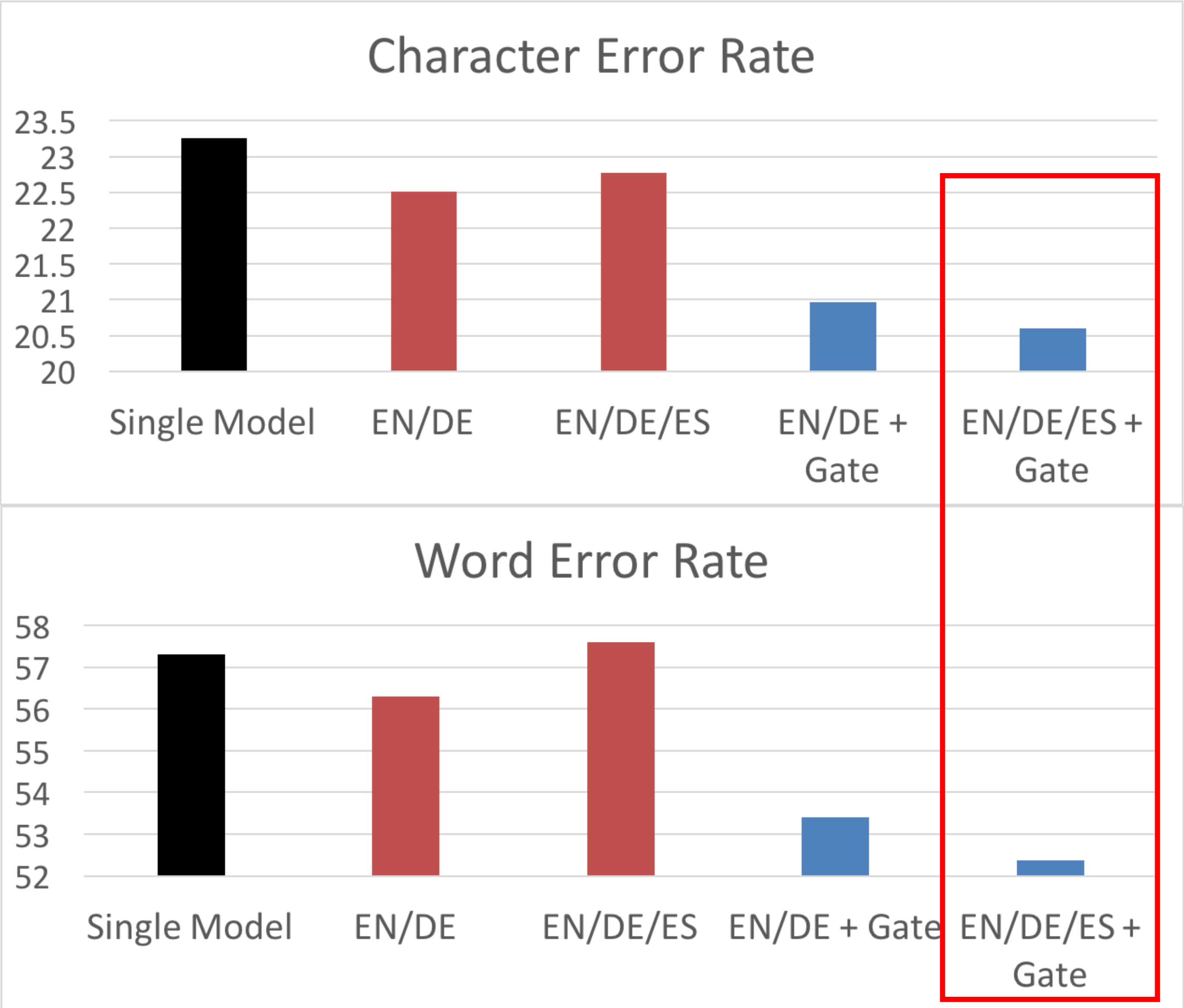


EN evaluation:
10.7% rel. impr. in CER, 7.0% rel. impr. in WER



- *without* Gate, no benefit increasing from 2 languages to 3 languages
- *with* Gate, additional gain increasing from 2 languages to 3 languages

DE evaluation:
11.4% rel. impr. in CER, and 8.6% rel. impr. in WER

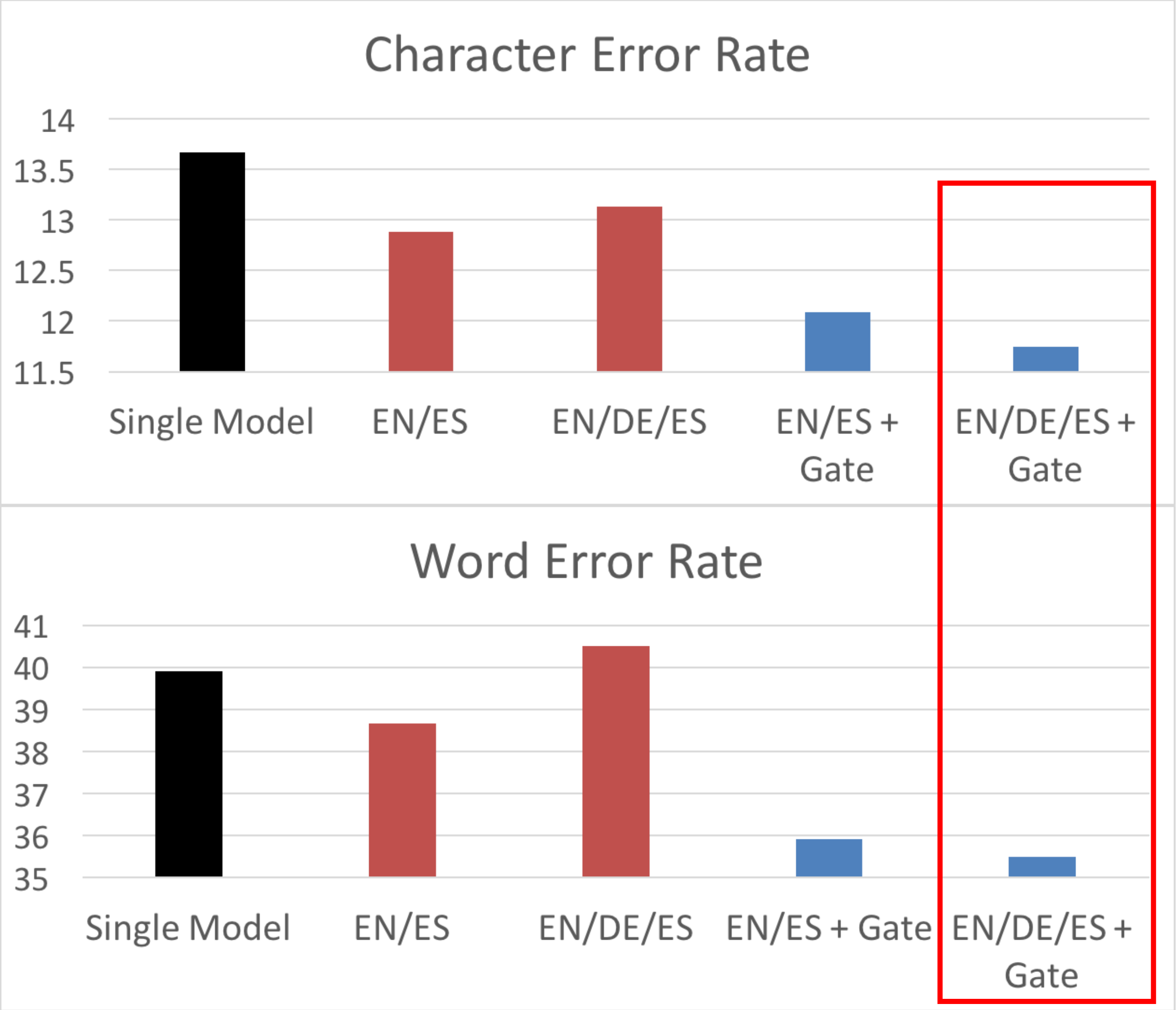


Character Error Rate

Word Error Rate

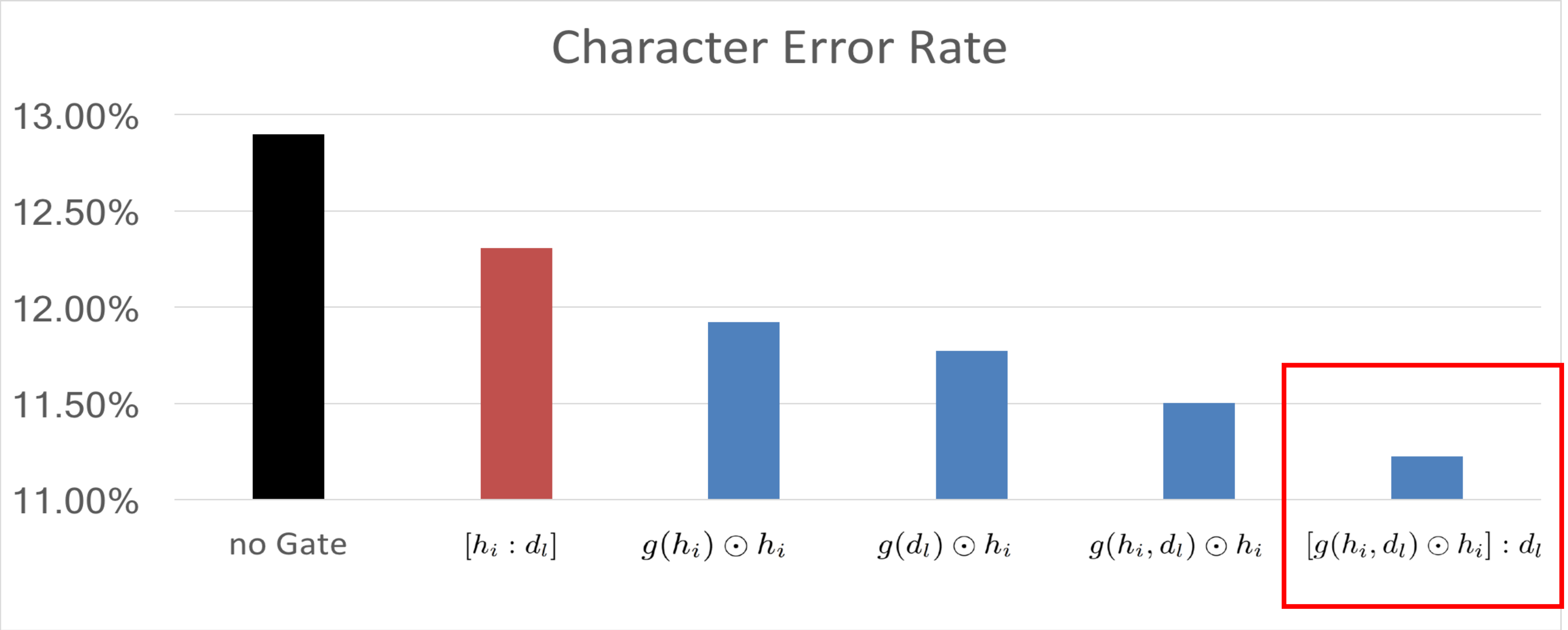
- *without* Gate, no benefit increasing from 2 languages to 3 languages
- *with* Gate, additional gain increasing from 2 languages to 3 languages

ES evaluation:
14.1% rel. impr. in CER, and 11.1% rel. impr. in WER



- *without* Gate, no benefit increasing from 2 languages to 3 languages
- *with* Gate, additional gain increasing from 2 languages to 3 languages

Different ways to add language information to the model



- Adding one-hot language ID input gives minimal improvement (+ 0.1M parameters)
- Proposed approach results in the largest improvement, (+ 0.4M parameters, much fewer than *cluster adaptive training*[1][2])

[1] *Li et al., multi-dialect speech recognition with a single sequence-to-sequence model, 2018*

[2] *Tan et al., cluster adaptive training for deep learning network based acoustic model, 2016*

Language-universal model can be a good initial model for creating a language-specific model

Initial Model	Fine Tune	DE CER (%)
--	DE (150h)	23.3
EN (1000h)	DE (150h)	21.4
EN + DE (300h)	DE (150h)	21.1
EN + ES + DE + gate (450h)	--	20.6
EN + ES + DE + gate (450h)	DE (150h)	**19.4**

- Fine-tuning DE from our universal model gets further gain - (5.8%)
- Our universal model is better initial model than EN (1000hr), well-trained monolingual from a different language - (9.3%)

Conclusions

- Our Language-Universal End-to-End Speech Recognition model
  - Does not require lexicon information and easy to maintain in production
  - Shows **7.0% - 11.1%** WER reduction over monolingual character-based model
  - Shows **9.1% - 12.4%** WER reduction over conventional MTL approach
  - Can be used as a **good initial model** for the further adaptation
    - Improves performance over bootstrapping from a well-trained monolingual from a different language
  - Need to evaluate with explicit language model