INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI, CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

東京工業大学
Tokyo Institute of Technology

Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning

Wenxin Hou [1], Yue Dong [1], Bairong Zhuang [1], Longfei Yang [1], Jiatong Shi [2], Takahiro Shinozaki [1]

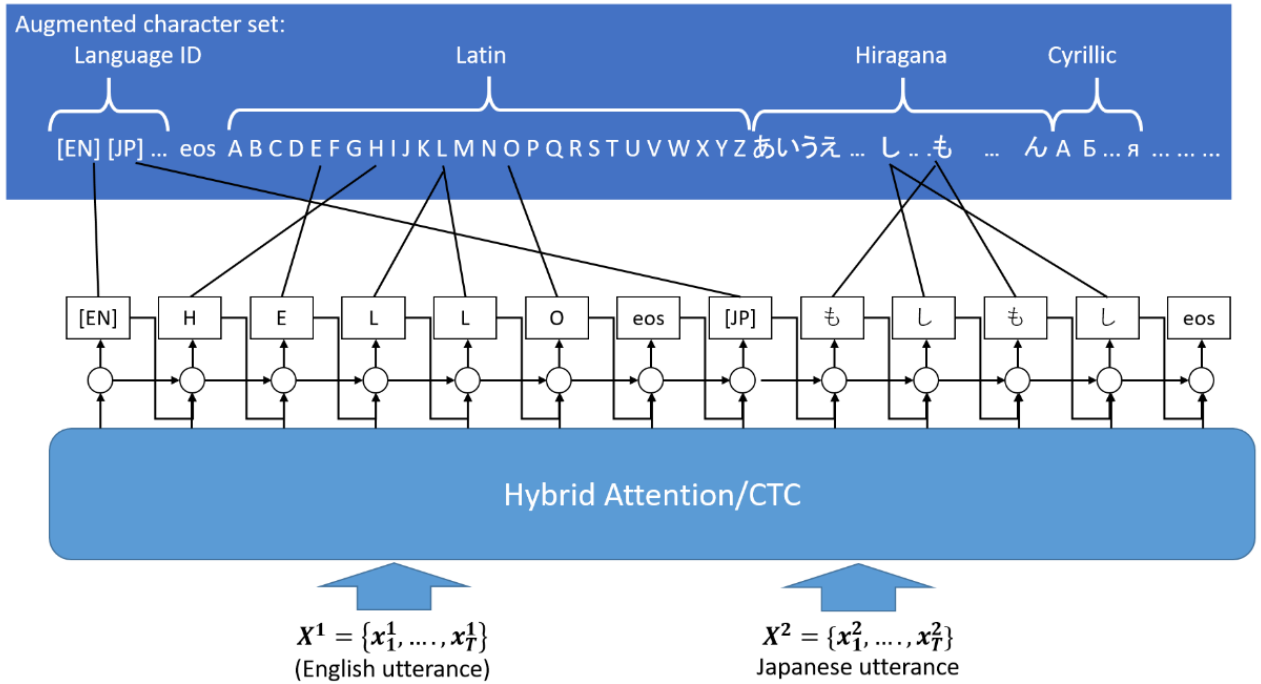
[1] Tokyo Institute of Technology
[2] Johns Hopkins University

http://www.ts.ip.titech.ac.jp/

- **Background**
  1. A single multilingual model saves efforts on deploying and maintaining in real-world application
  2. A pretrained multilingual ASR model can benefit low-resource scenarios
  3. End-to-end ASR on a large number of languages and large amount of data is rarely explored so far

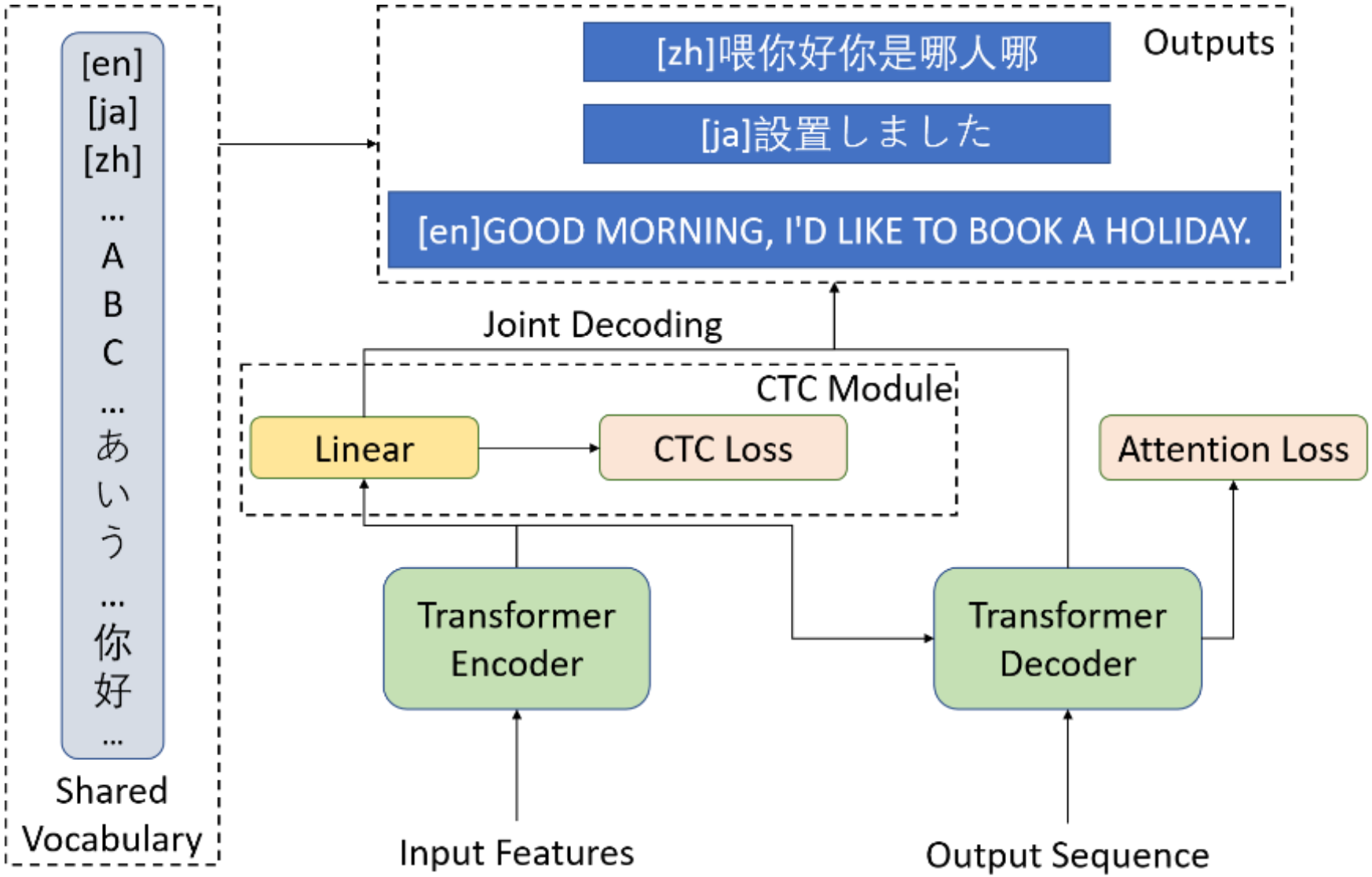
- **Motivation**
  1. Investigate into very large-scale multilingual ASR with automatic language identification based on hybrid CTC/Attention Transformer architecture
  2. Discuss the transfer learning performance to low-resource languages with the pre-trained model

- Watanabe et al. [1] first proposed the E2E language-independent architecture for joint ASR and language identification(LID) with Hybrid CTC/Attention architecture on 10 languages



- Cho et al. [2] showed that transfer learning from 10 languages could improve performance on 4 low-resource languages

[1] Watanabe, et al, “Language independent end-to-end architecture for joint language identification and speech recognition”, ASRU 2017, pp. 265–271
[2] J. Cho, et al, “Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling”, IEEE SLT 2018, pp. 521–527


[en]
[ja]
[zh]
...
A
B
C
...
あ
い
う
...
你
好
...
Shared Vocabulary
[zh]喂你好你是哪人哪
[ja]設置しました
[en]GOOD MORNING, I'D LIKE TO BOOK A HOLIDAY.
Outputs
Joint Decoding
CTC Module
Linear
CTC Loss
Attention Loss
Transformer Encoder
Transformer Decoder
Input Features
Output Sequence

- **Hybrid CTC/Attention Architecture**

  1. Multi-Task Learning:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{ctc}} + (1 - \alpha) \mathcal{L}_{\text{att}}$$

  2. Joint Decoding:

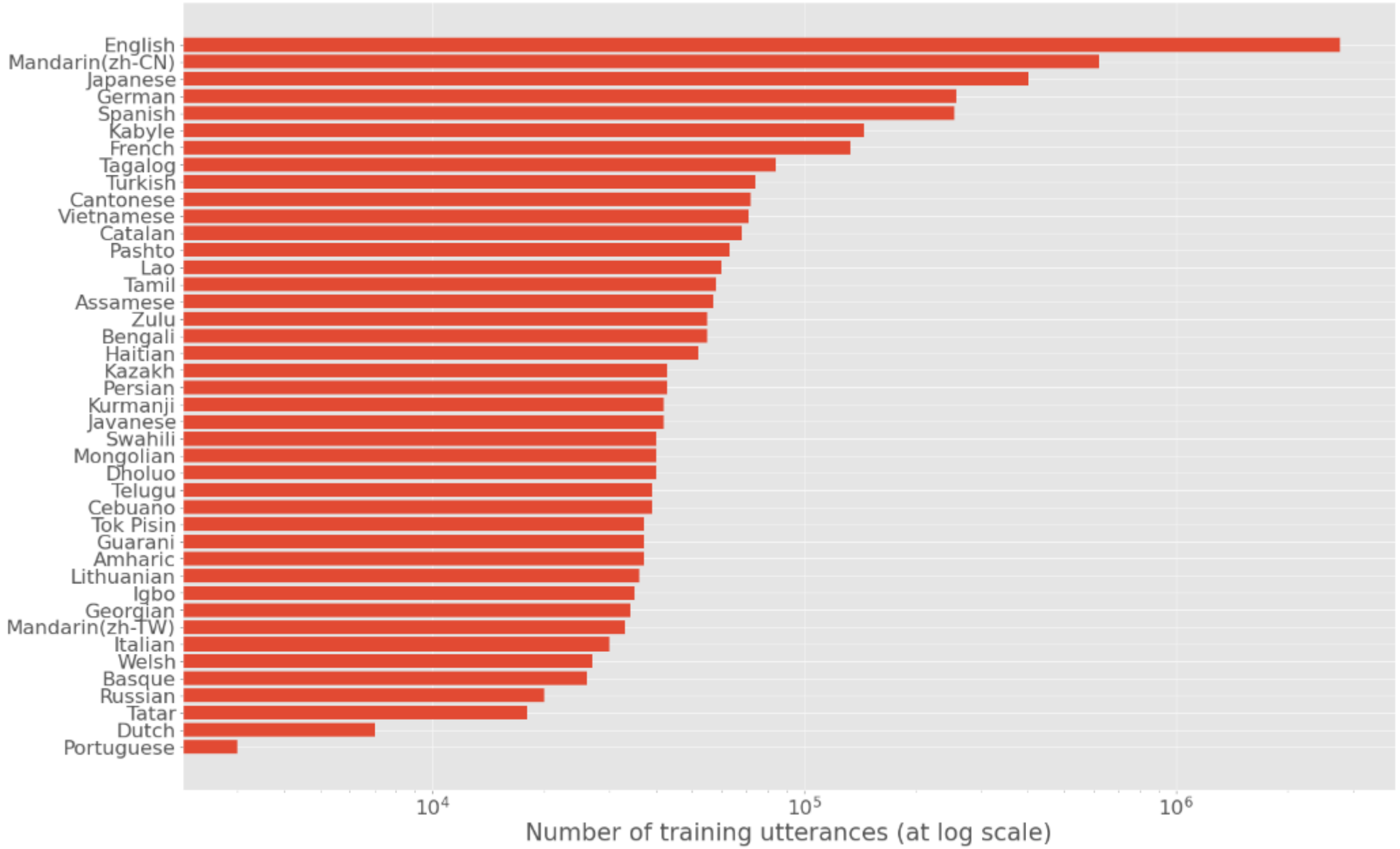
$$\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \{ \lambda \log P_{\text{ctc}}(Y|X) + (1-\lambda) \log P_{\text{att}}(Y|X) \}$$

- **Language-Independent Architecture**

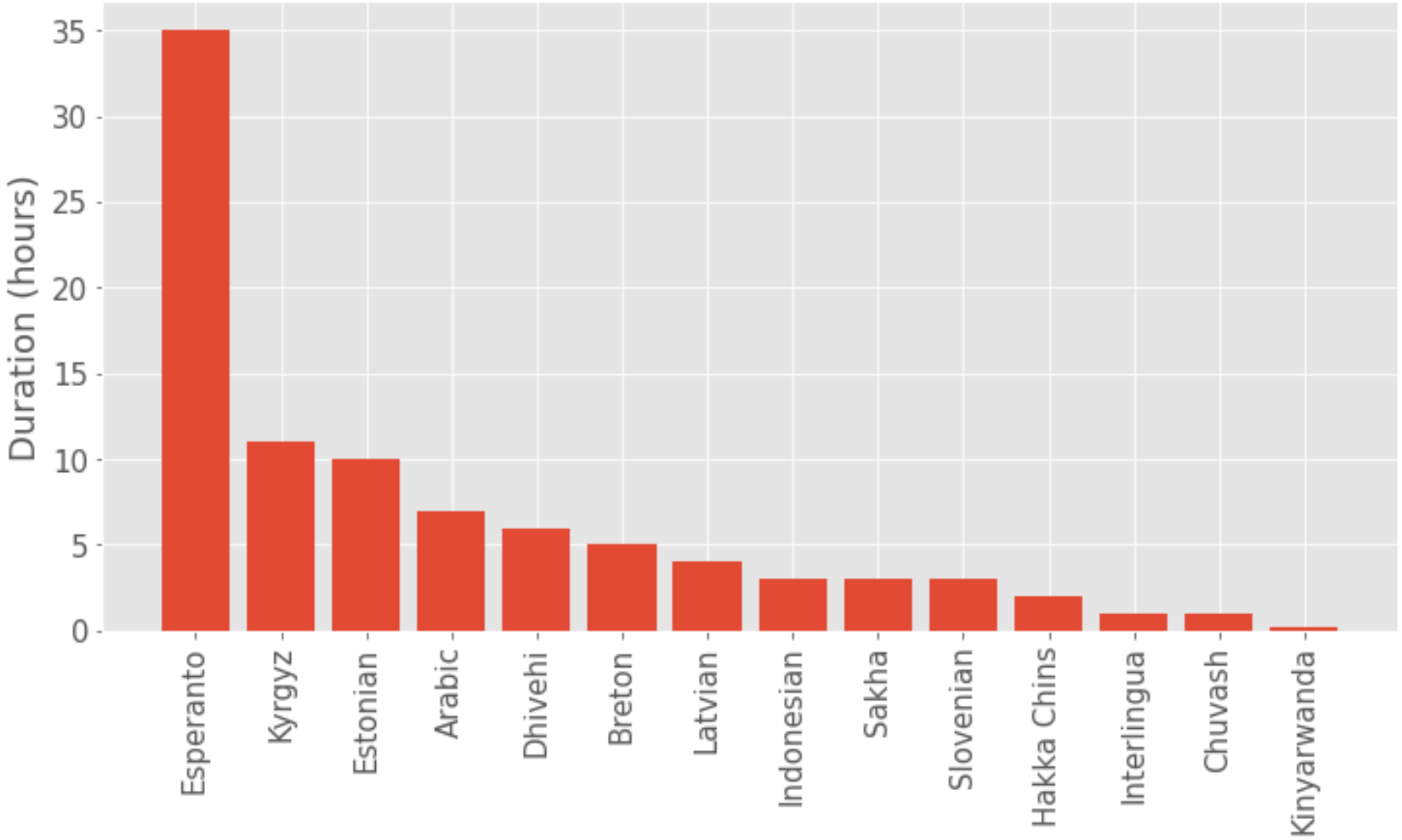
  1. Share modeling units (characters, subwords) of all languages in one vocabulary
  2. Prepend corresponding language IDs (e.g. [en]) to the beginning of utterance labels

- **Data Set**
  - 42 languages, 5000 hours, 6 million utterances from 11 corpora

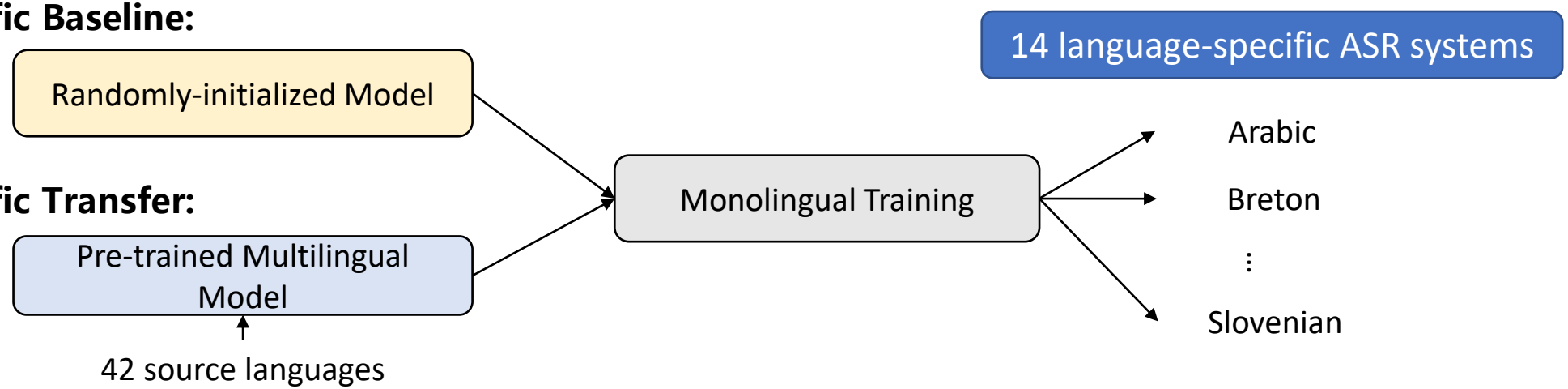
  - 14 low-resource languages from Common Voice database
- **Modeling Units**: 7,381 characters (Char.) / 15,943 subwords (SubW.) + 60 non-language symbols
- **Evaluation Metrics**:
  - Character Error Rate (CER)
  - Word Error Rate (WER)
  - Language Identification (LID) Accuracy


English
Mandarin(zh-CN)
Japanese
German
Spanish
Kabyle
French
Tagalog
Turkish
Cantonese
Vietnamese
Catalan
Pashto
Lao
Tamil
Assamese
Zulu
Bengali
Haitian
Kazakh
Persian
Kurmanji
Javanese
Swahili
Mongolian
Dholuo
Telugu
Cebuano
Tok Pisin
Guarani
Amharic
Lithuanian
Igbo
Georgian
Mandarin(zh-TW)
Italian
Welsh
Basque
Russian
Tatar
Dutch
Portuguese
10^4
10^5
10^6
Number of training utterances (at log scale)

- Randomly split 80% utterances for training and 20% for testing

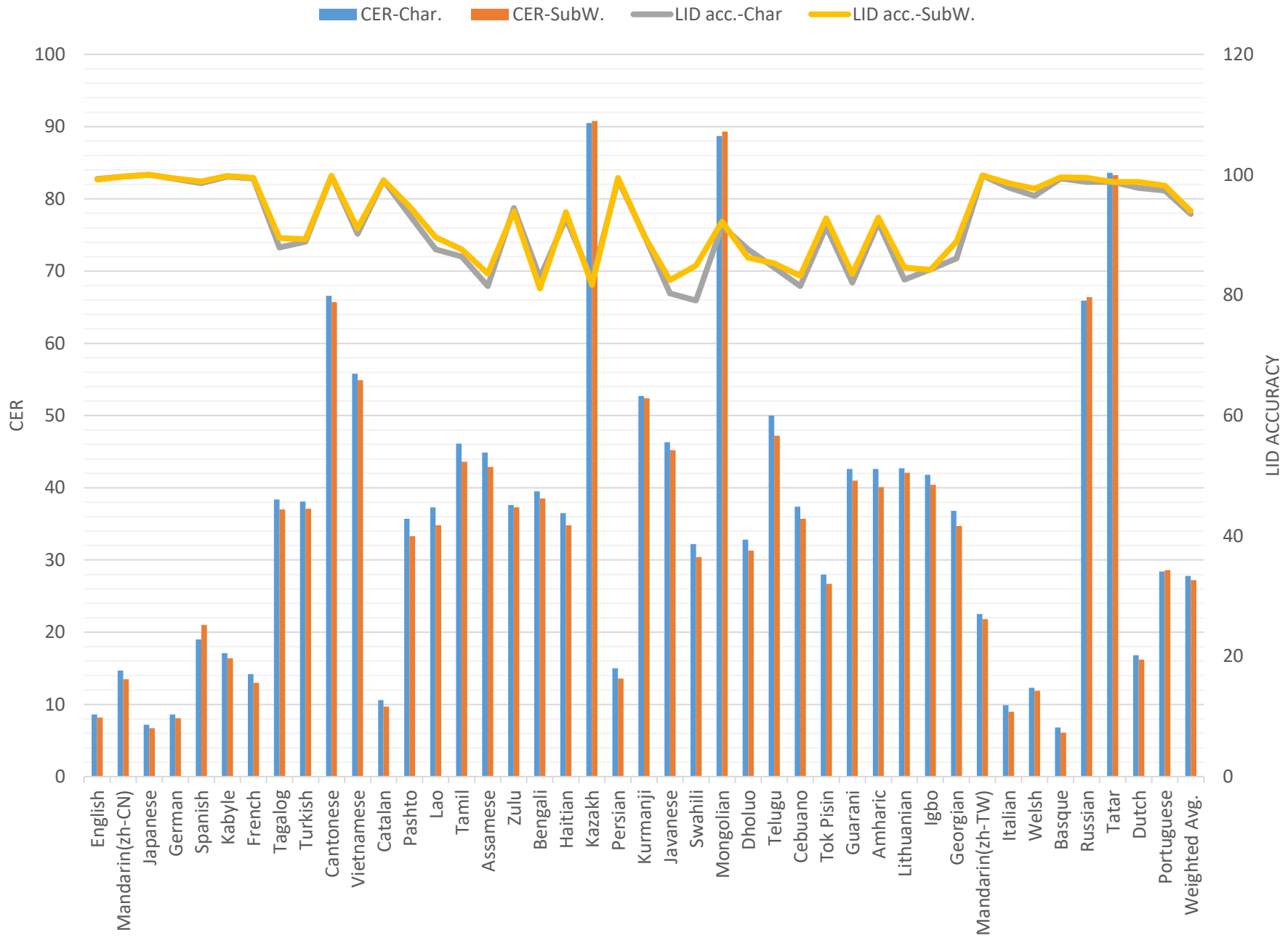


- Fine-tune the network by replacing output layers of CTC & Transformer decoder
- Compare transfer performance under language specific & independent settings
  - **Language-specific Baseline:**
  - **Language-specific Transfer:**

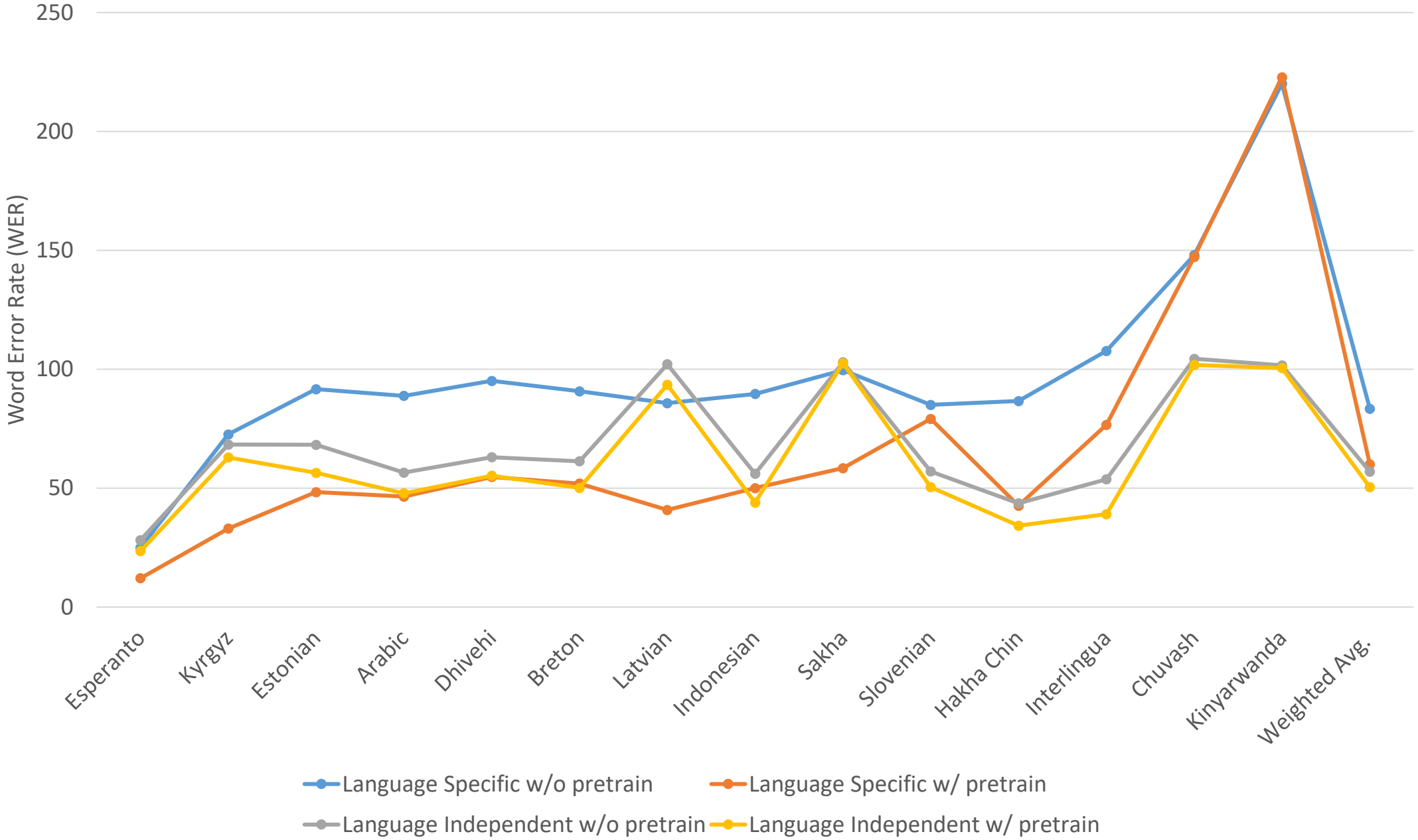


  - **Language-independent Baseline:**
  - **Language-independent Transfer:**

03 **Experiments** | ASR Results (CER&LID) on 42 Languages

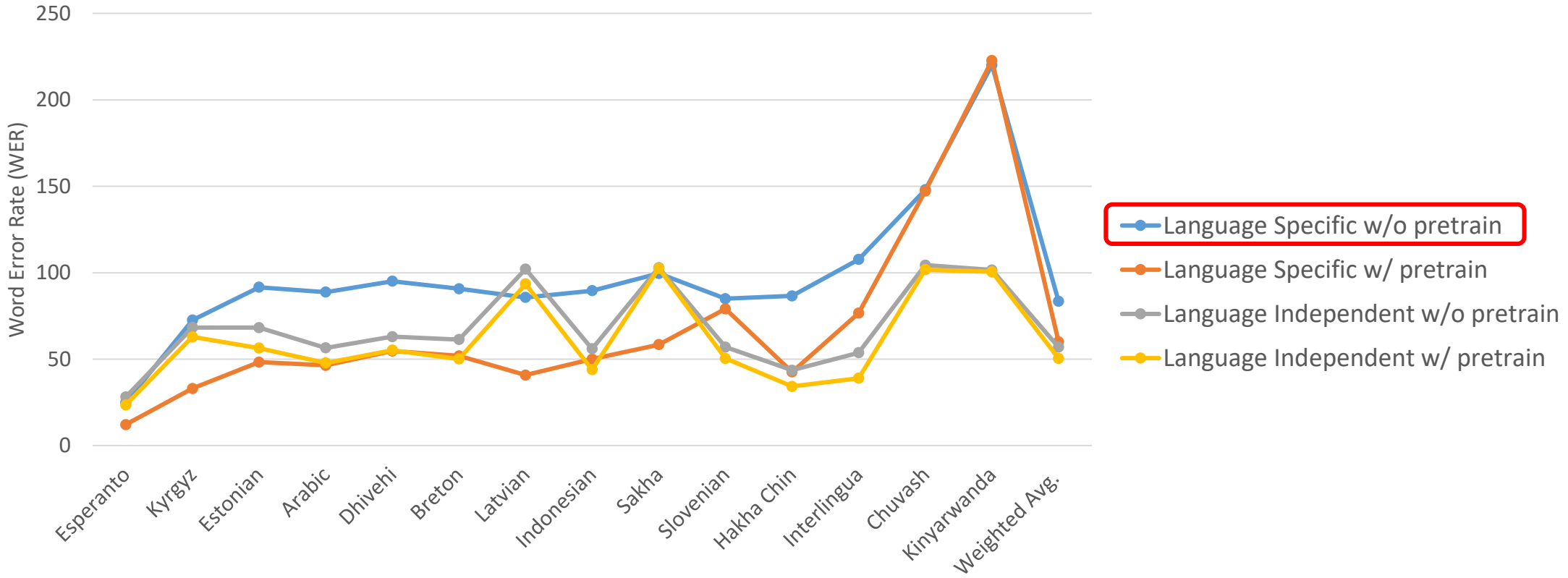


03 **Experiments** | ASR Results (WER) on 14 Low-resource Languages



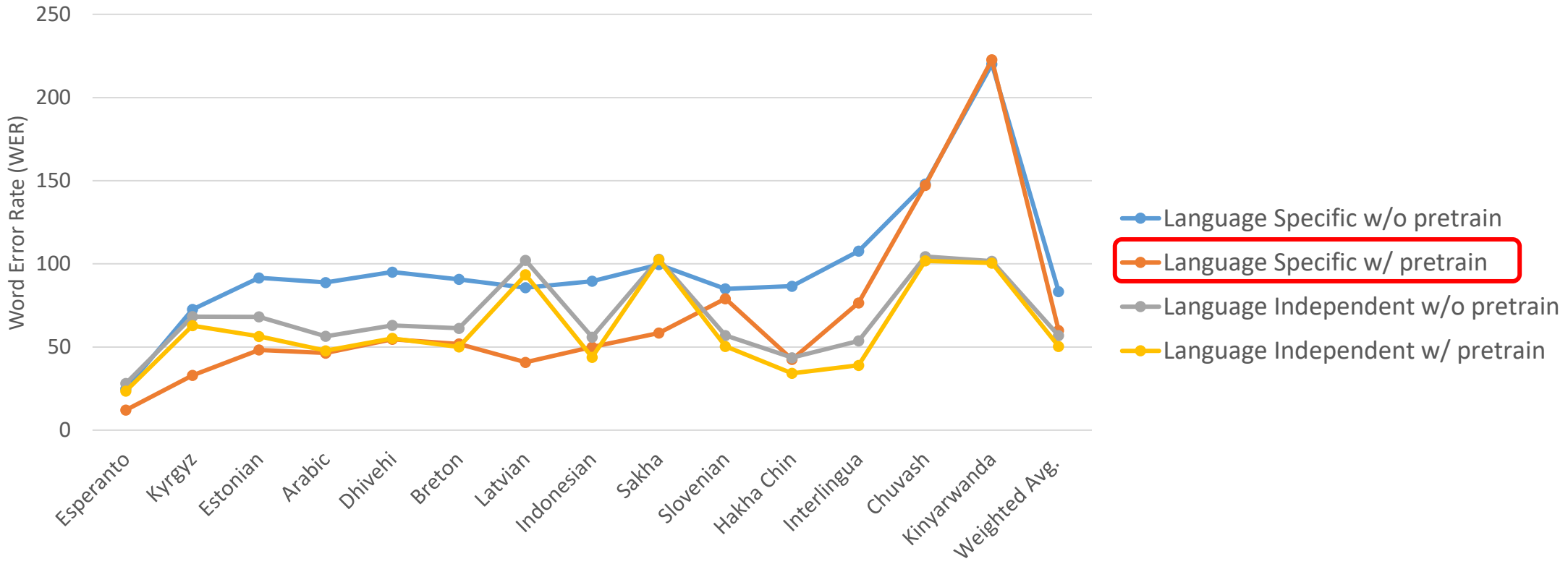
- **Language-specific baseline (blue)**

1. Relatively high WER due to lack of training data
2. WER goes beyond 100% on Interlingua, Chuvash & Kinyarwanda (all <= 1 hour)
3. 24.9% on Esperanto (35 hours)



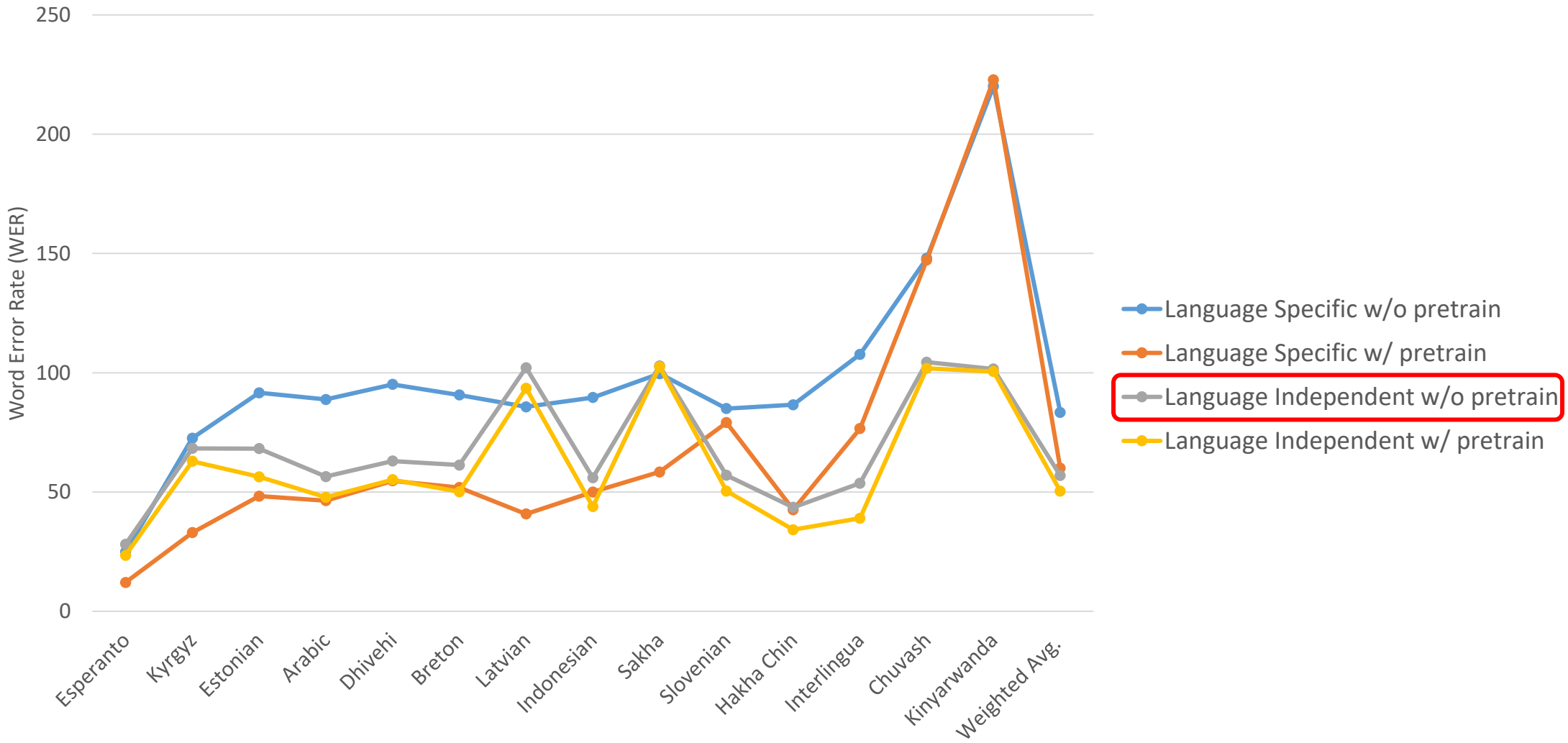
- **Language-specific transfer (red)**

1. Most languages witness a significant drop in WER, indicating the effectiveness of pre-training
2. However, WER conversely increases on Kinyarwanda due to extremely low resource (0.25 hours)
3. Weighted average WER is reduced by 28.1% compared with language-specific baseline



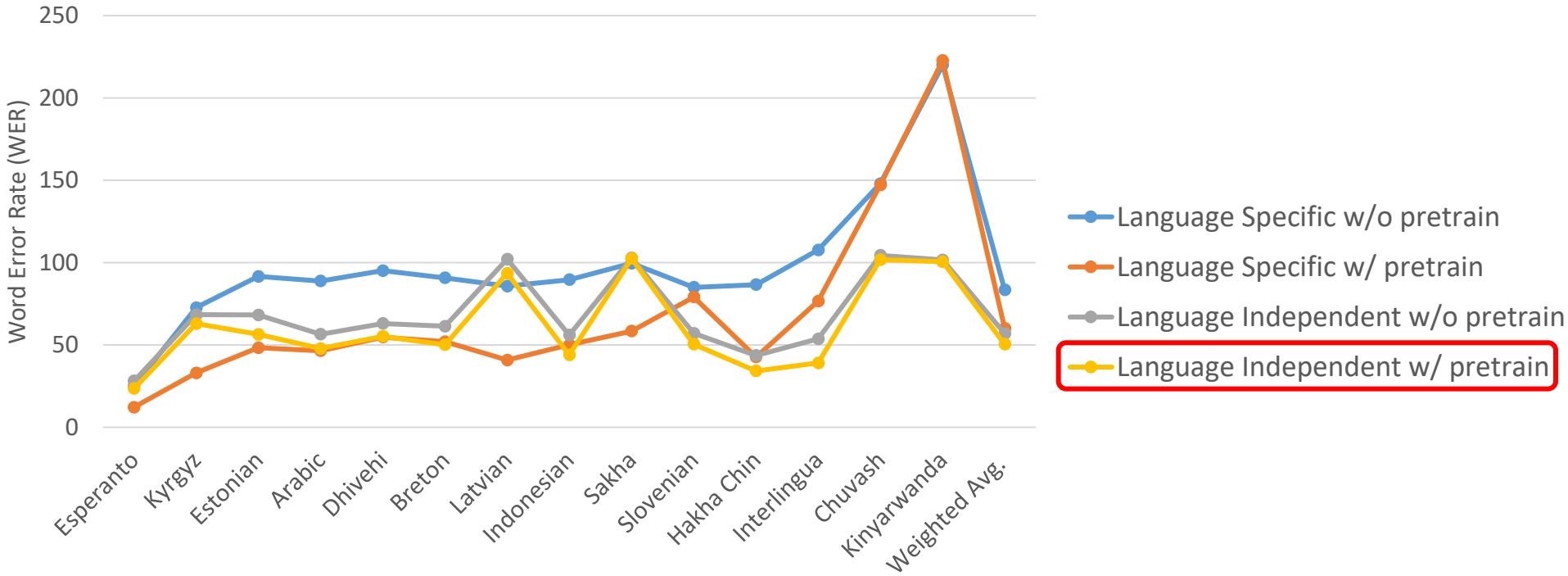
- **Language-independent baseline (grey)**

1. Significant reduction in WER compared with language-specific baselines



- **Language-independent transfer (yellow)**

1. WER stably decreases on all languages, including Kinyarwanda
2. Side effects may occur (e.g. Arabic), resulting in slightly worse performance compared with language-specific transfer
3. Weighted average WER is reduced by 11.4% compared with language-independent baseline

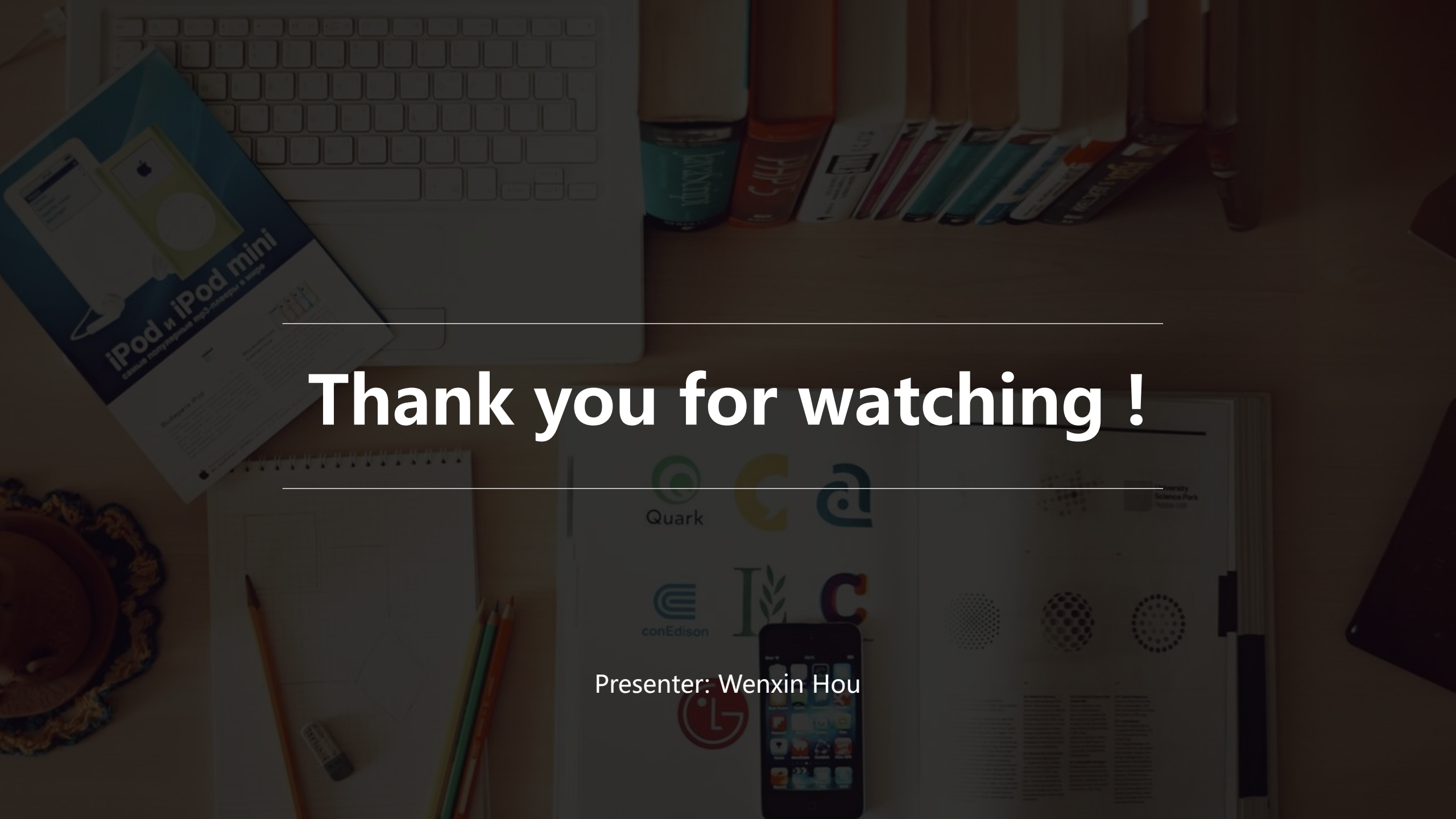


04 Conclusions

- The presented large-scale multilingual model trained on 42 languages shows promising results in terms of average CER

- WER performances on low-resource languages can be greatly improved by applying large-scale multilingual pre-training

- Future work includes investigating into
  1. Leveraging linguistic similarities between languages to further improve multilingual ASR and LID
  2. Efficient data sampling methods to handle data imbalance problem



Thank you for watching !

Presenter: Wenxin Hou