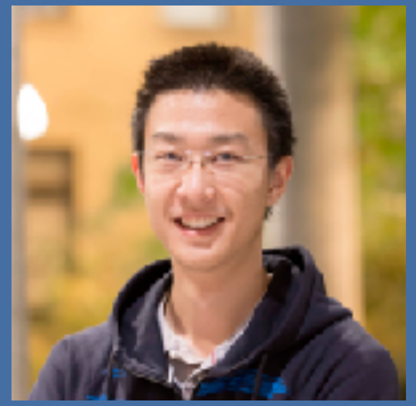
Self-supervised for speech processing

Facebook AI Research


Alexei Baevski


Alexis Conneau


Steffen Schneider


Henry Zhou


Abdelrahman Mohamed


Anuroop Sriram


Naman Goyal


Wei-Ning Hsu


Michael Auli


Kritika Singh


Yatharth Saraf


Geoffrey Zweig


Qiantong Xu

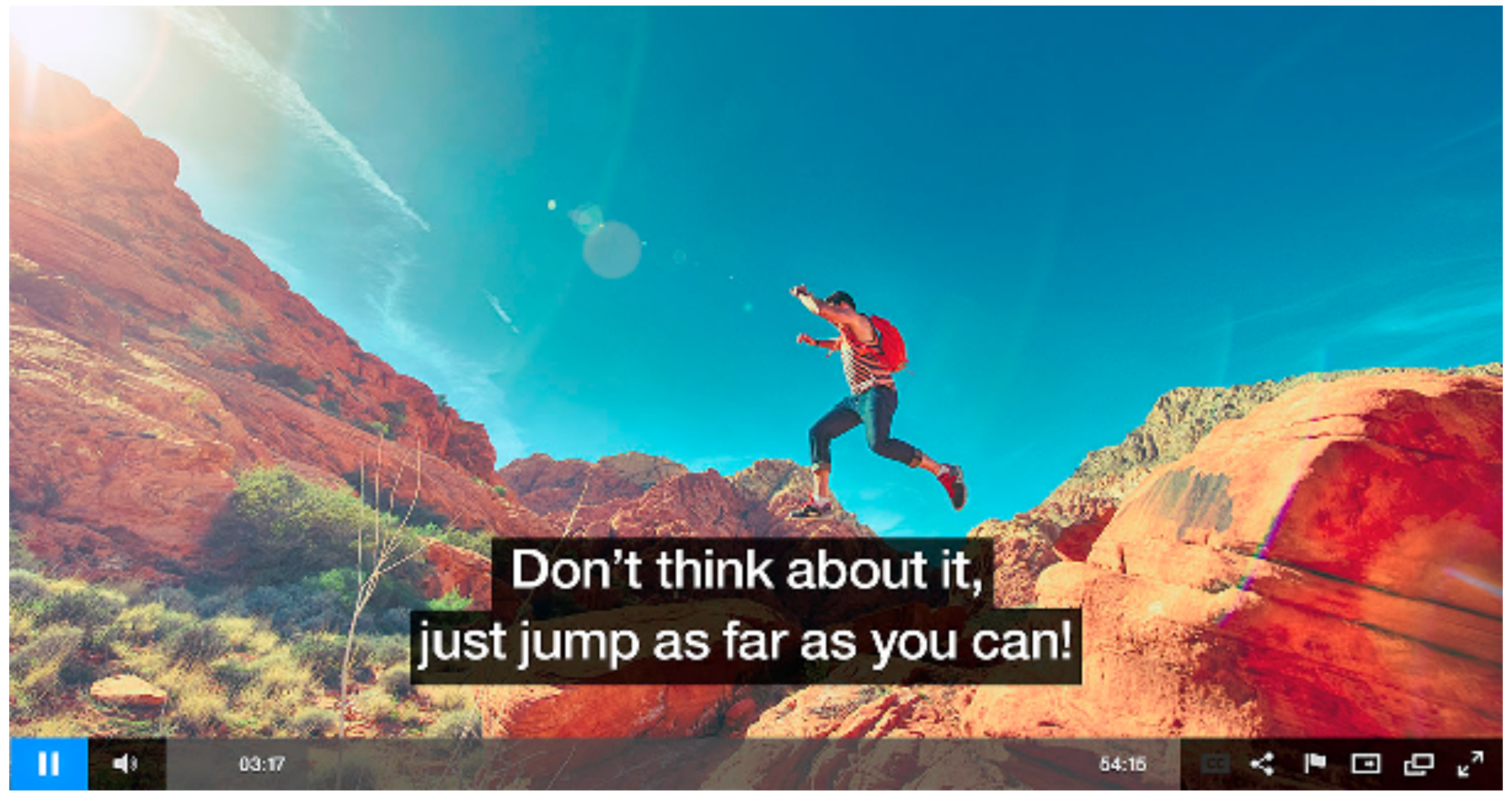

Tatiana Likhomanenko


Paden Tomasello


Ronan Collobert


Gabriel Synnaeve

Speech technology



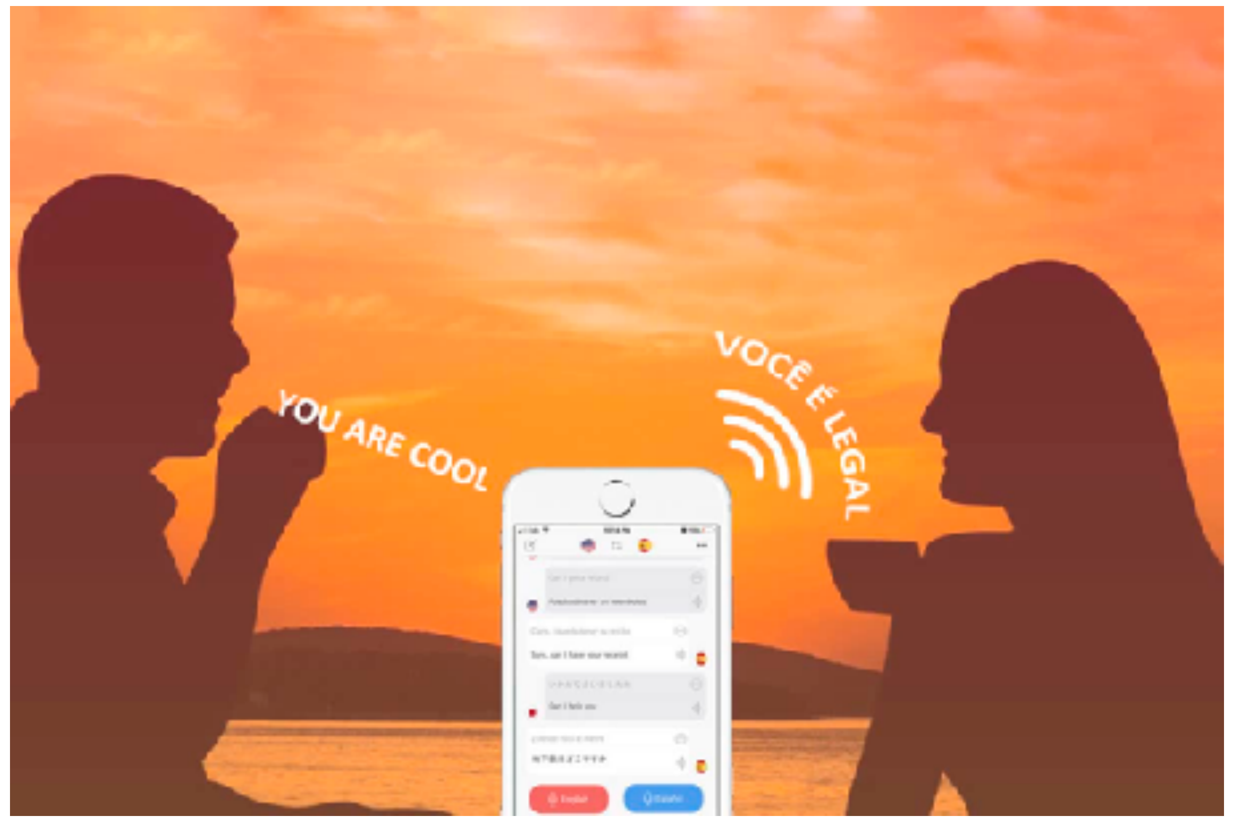
Video captioning


Mobile devices


Home devices

Speech applications

- **Speech to text (Speech recognition)**
- Text to speech
- Keyword spotting ("Hey Alexa/Portal")
- Speaker identification
- Language identification
- Speech translation

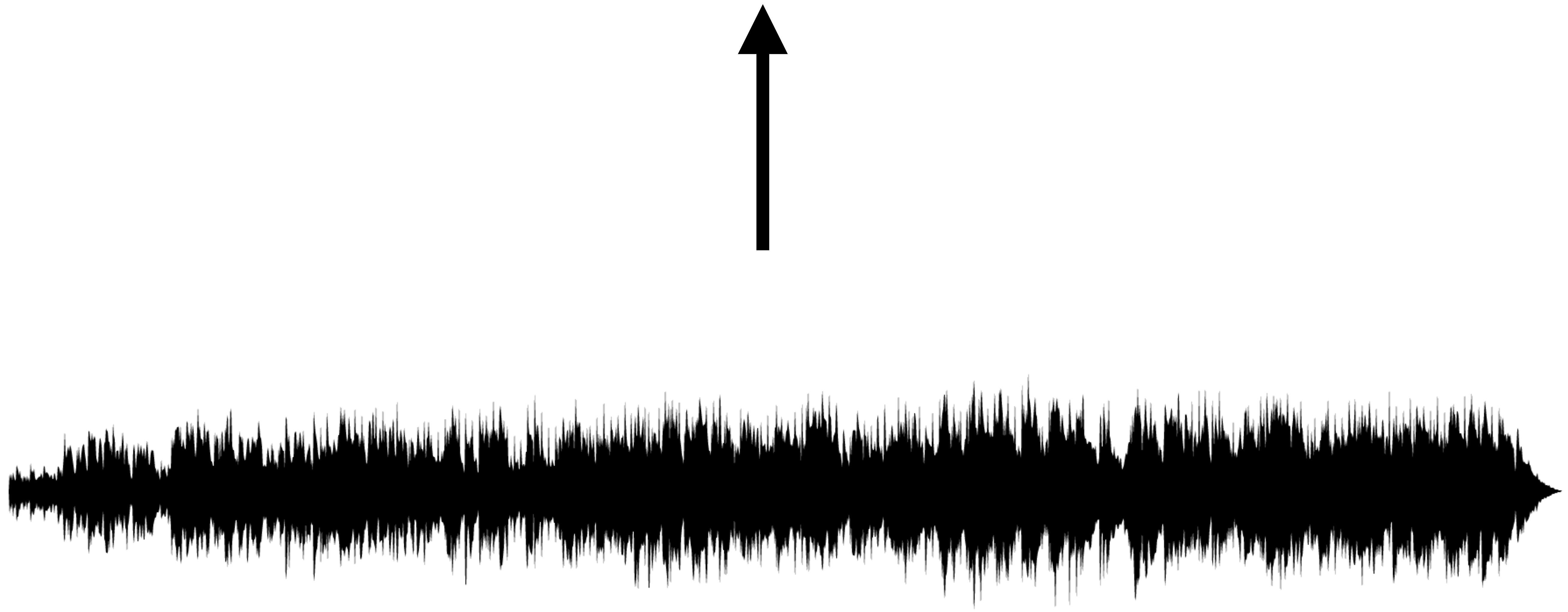


Overview

- Speech recognition
- Speech processing with less supervision / self-supervised learning
- Cross-lingual self-supervised learning for speech

Speech recognition

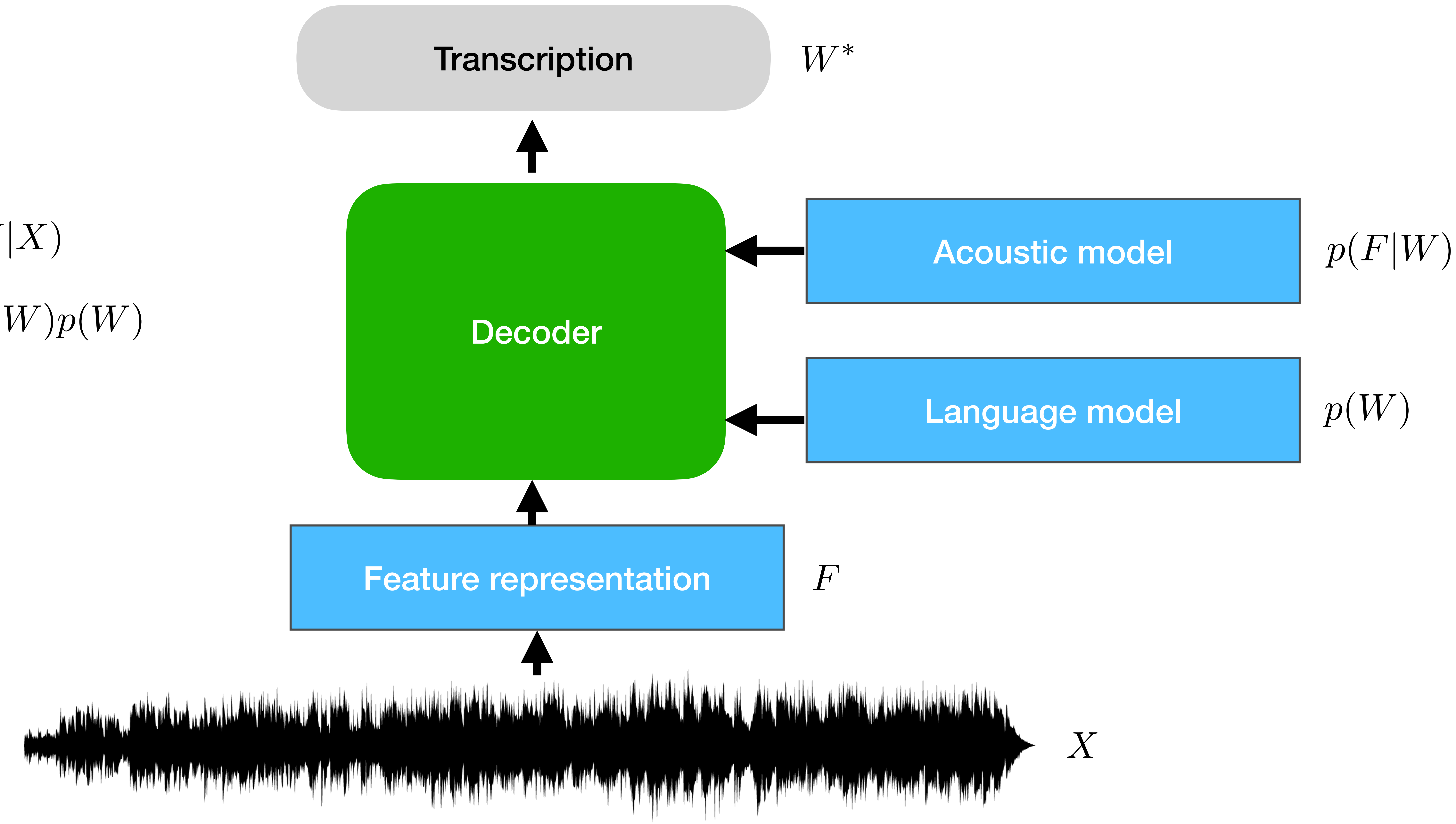
I like black tea with milk



Traditional automatic speech recognition (ASR)

$$W^* = \arg\max_W p(W|X)$$

$$W^* = \arg\max_W p(F|W)p(W)$$



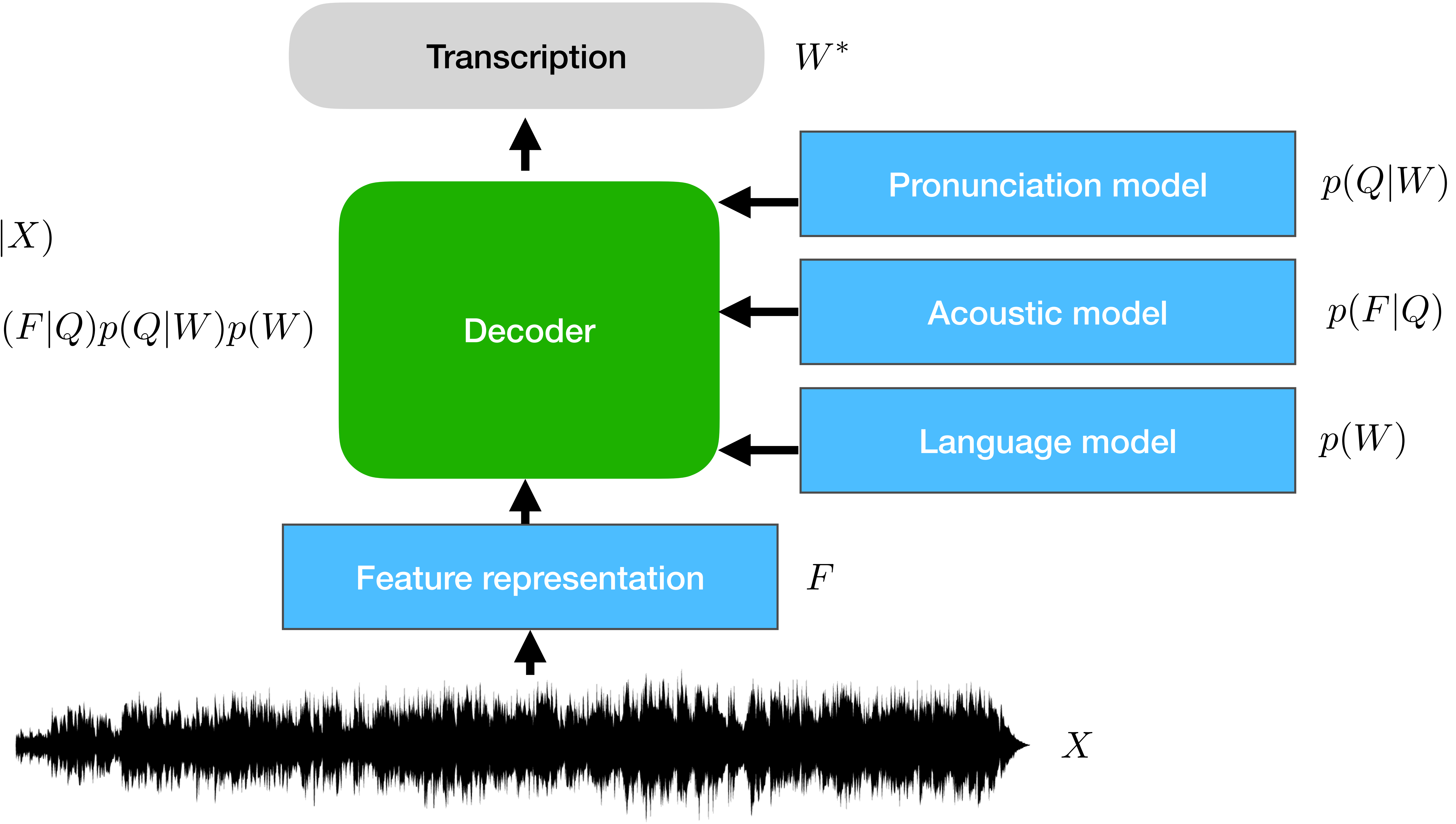
Traditional automatic speech recognition (ASR)

- Represent words as sequences of phonemes
- hello = h eh l ow
- Distinct units of sound to distinguish words

Traditional automatic speech recognition (ASR)

$$W^* = \arg\max_W p(W|X)$$

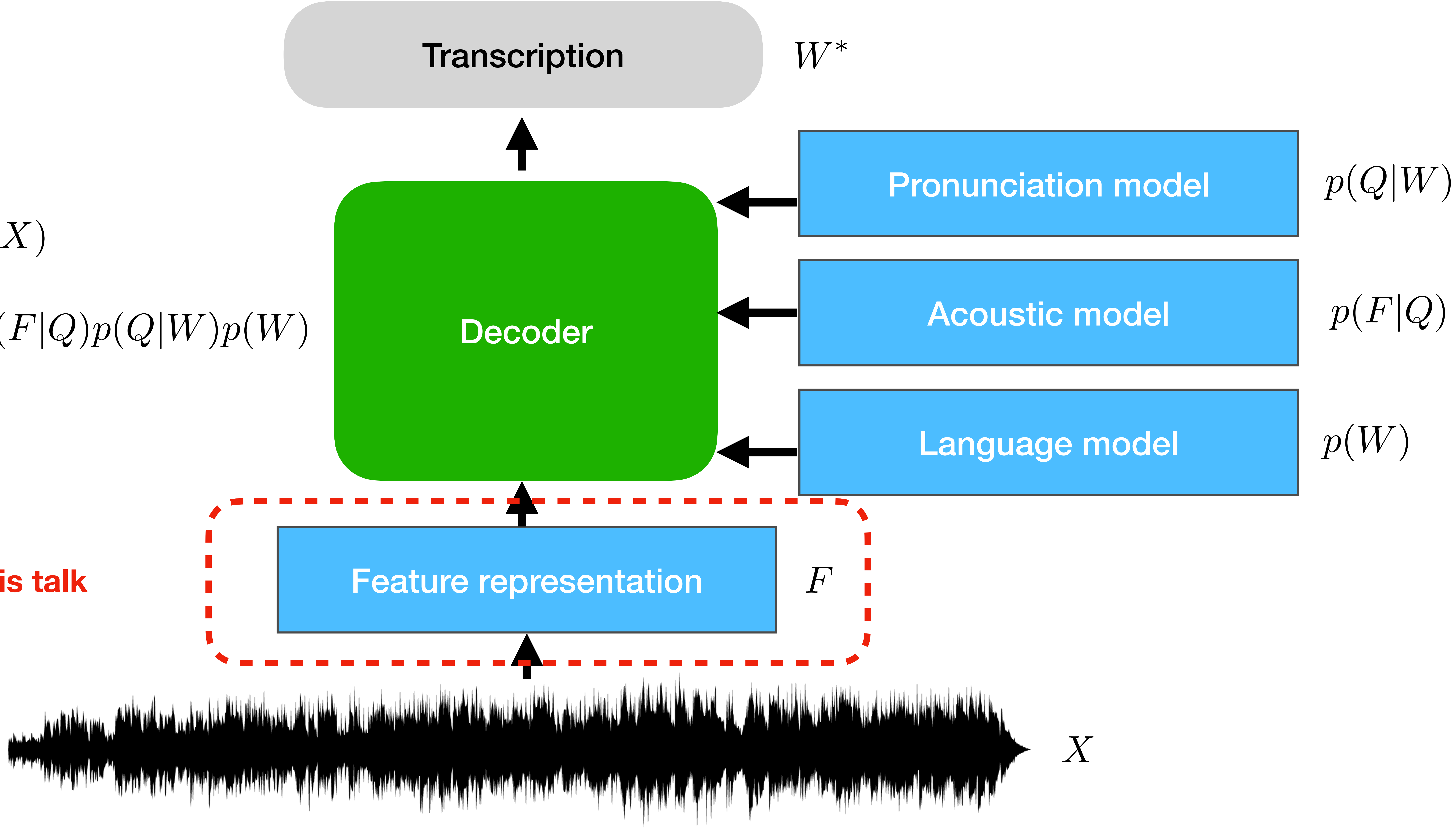
$$W^* = \arg\max_W \sum_Q p(F|Q)p(Q|W)p(W)$$



Traditional automatic speech recognition (ASR)

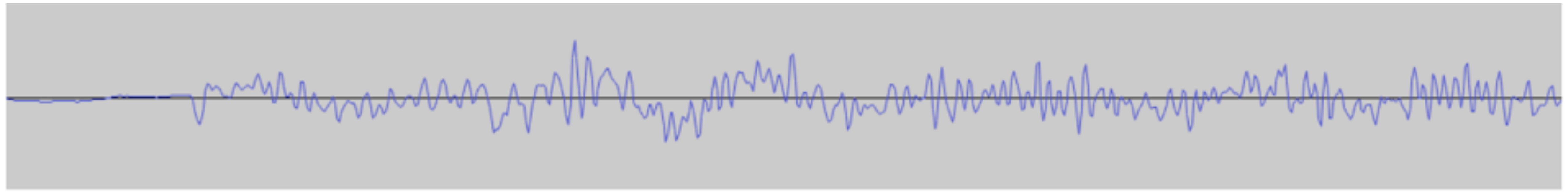
$$W^* = \arg\max_W p(W|X)$$

$$W^* = \arg\max_W \sum_Q p(F|Q)p(Q|W)p(W)$$



Focus of this talk

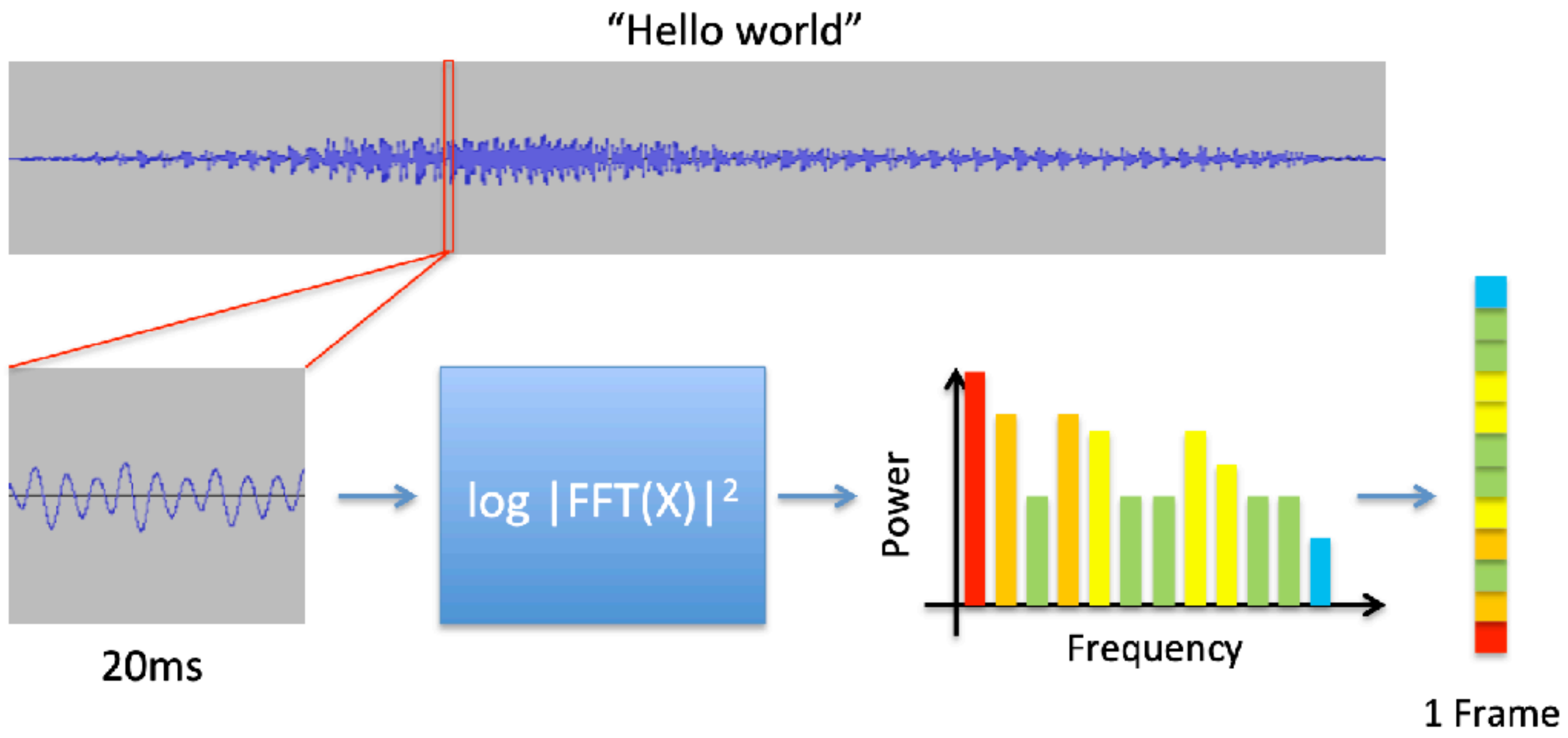
Feature representation



- Typical sample rates for speech: 8KHz, 16KHz.
- Traditionally: build spectrogram

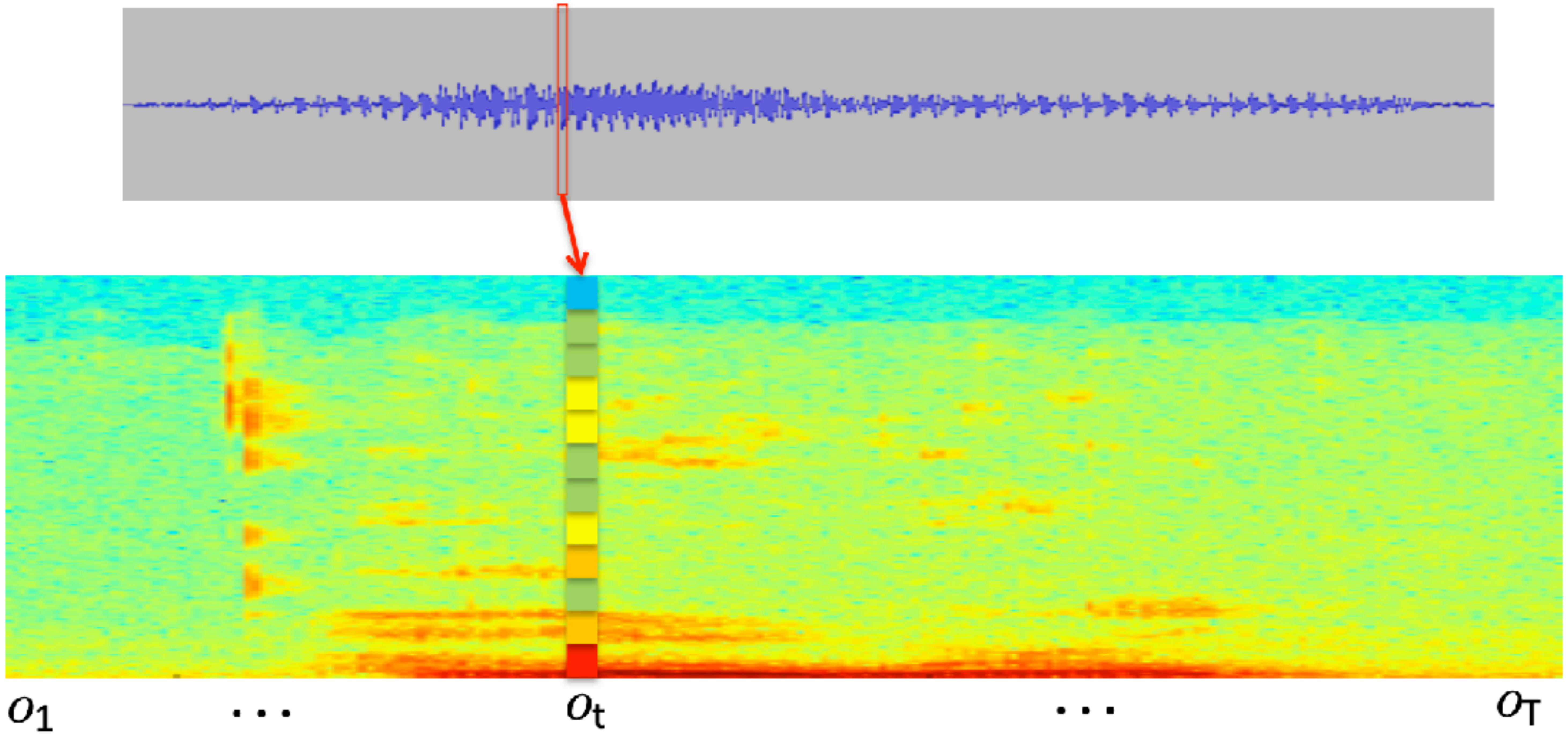
Spectrogram

- Small window, e.g., 20ms of waveform
  - Compute FFT and take magnitude
  - Describes frequency content in local window



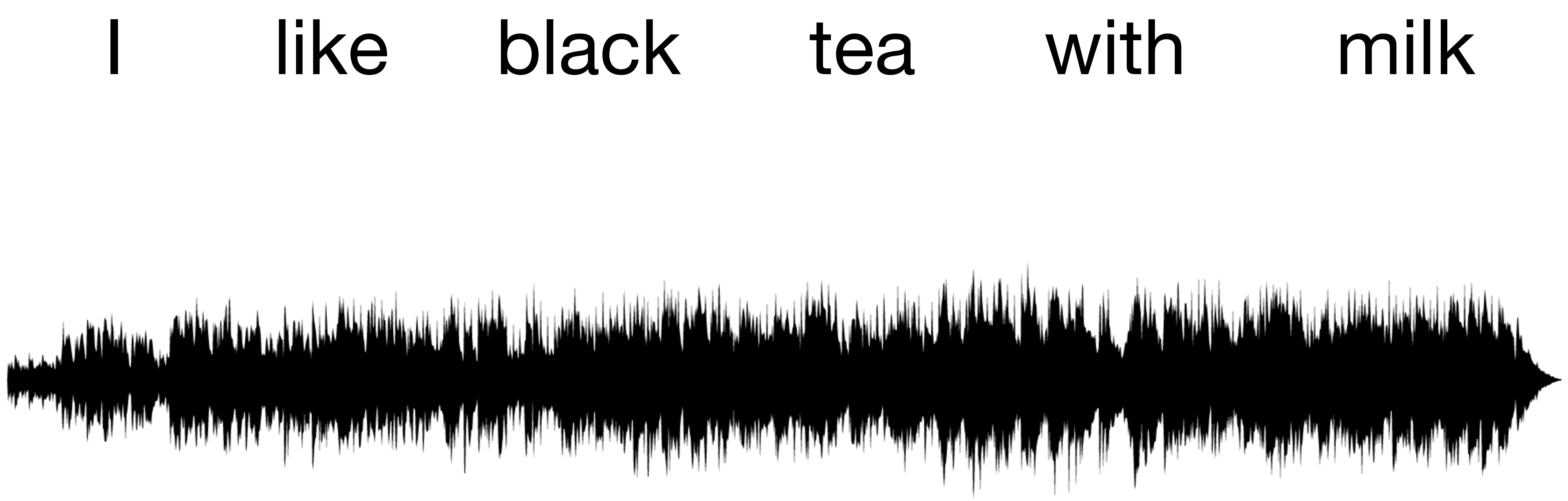
Spectrogram

- Concatenate frames from adjacent windows to form a spectrogram



Self-supervised speech representation learning

Training speech recognition models



- Train on 1,000s of hours of transcribed data for good systems.
- Many languages, dialects, domains etc.

Supervised Machine learning

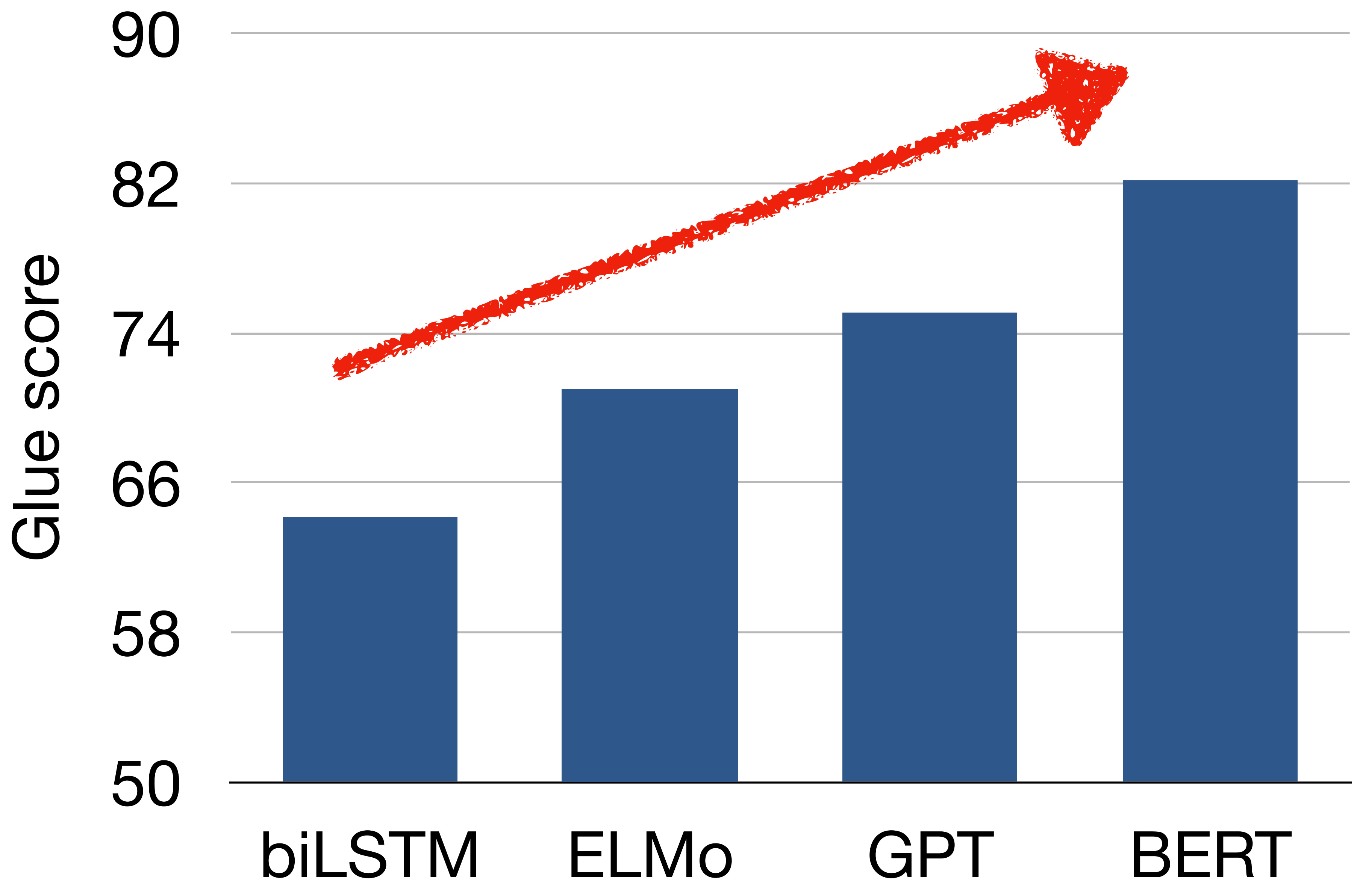
( , **cat**)

potential train/test mismatch 

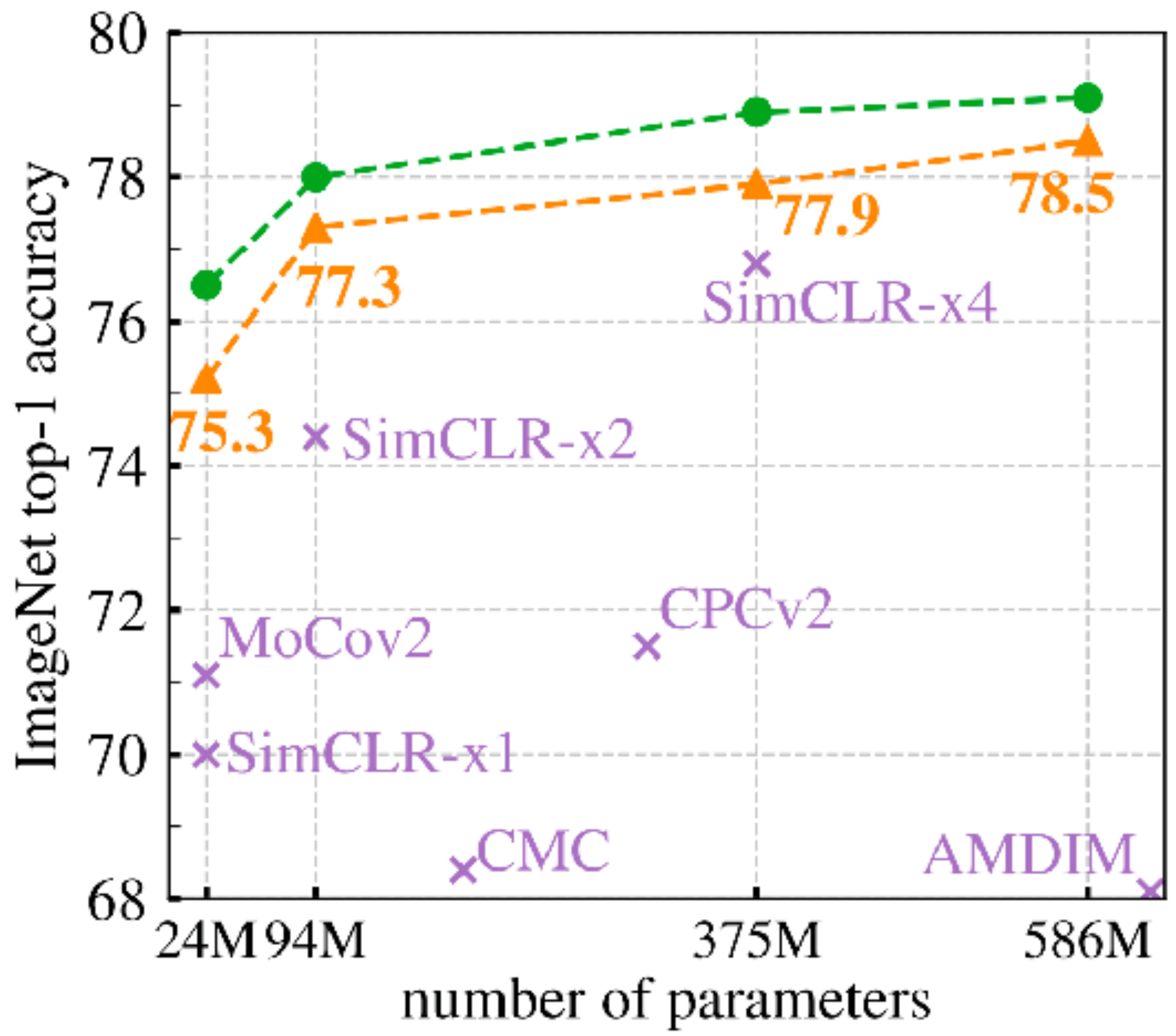
Need to annotate lots of data!

Meanwhile in other fields

Pre-training in NLP



Pre-training in Computer Vision

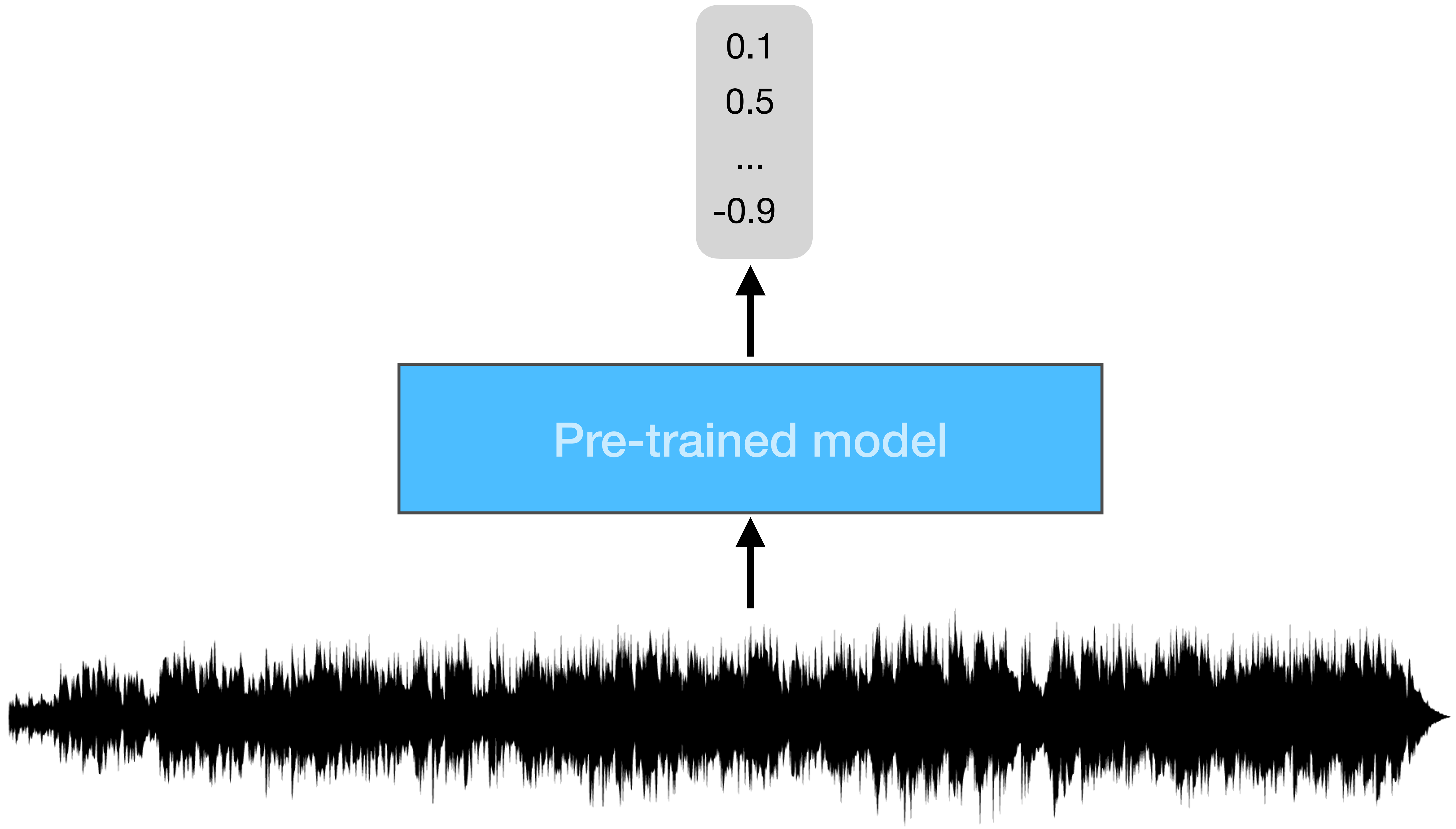


Thanks to Priya Goyal for sharing the vision graph.

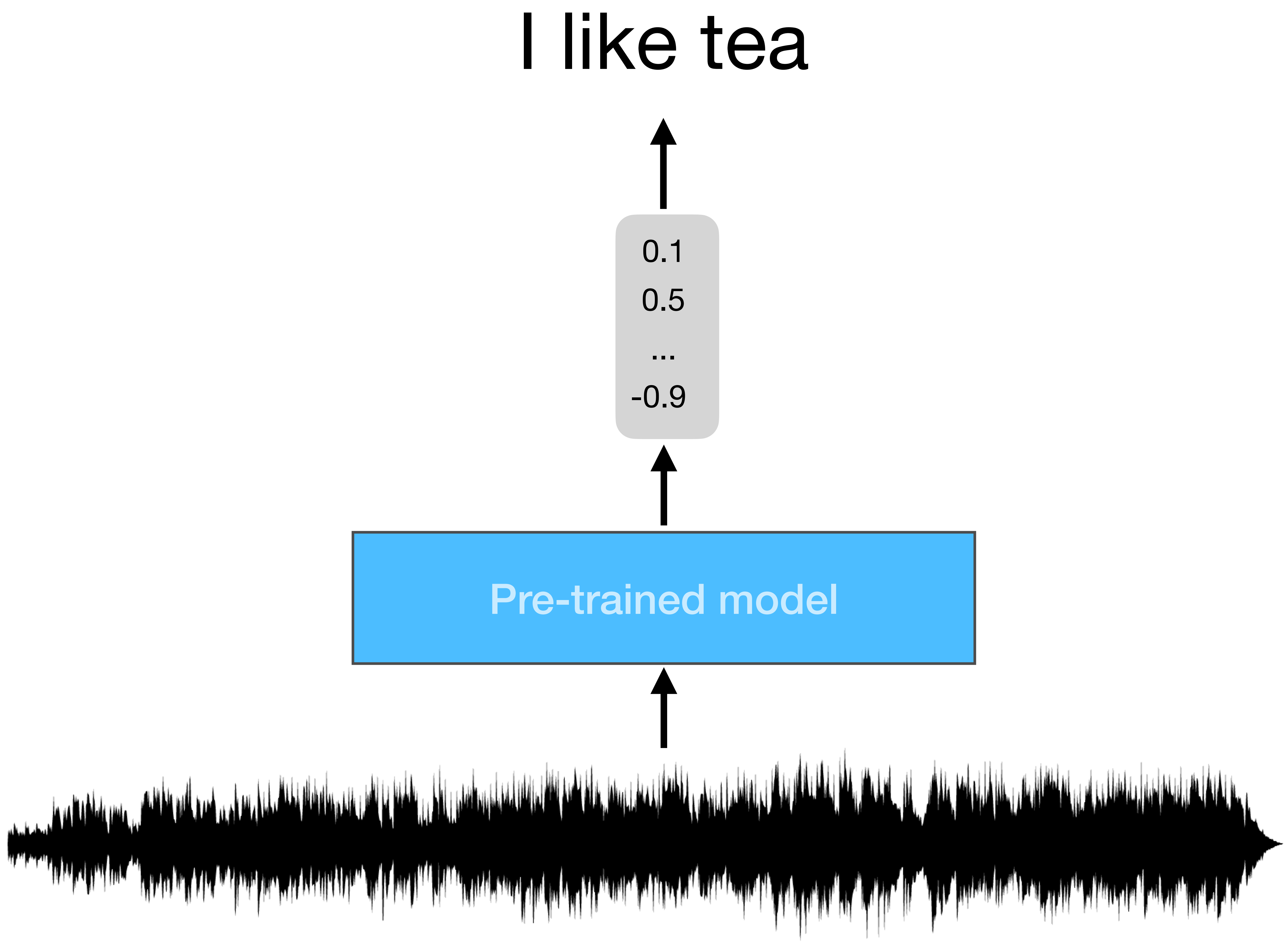
Unsupervised / Self-supervised Pre-training

- Learn good representations **without labels**
- NLP: Predict occluded parts of sentence
- Vision: make representations invariant to augmentations

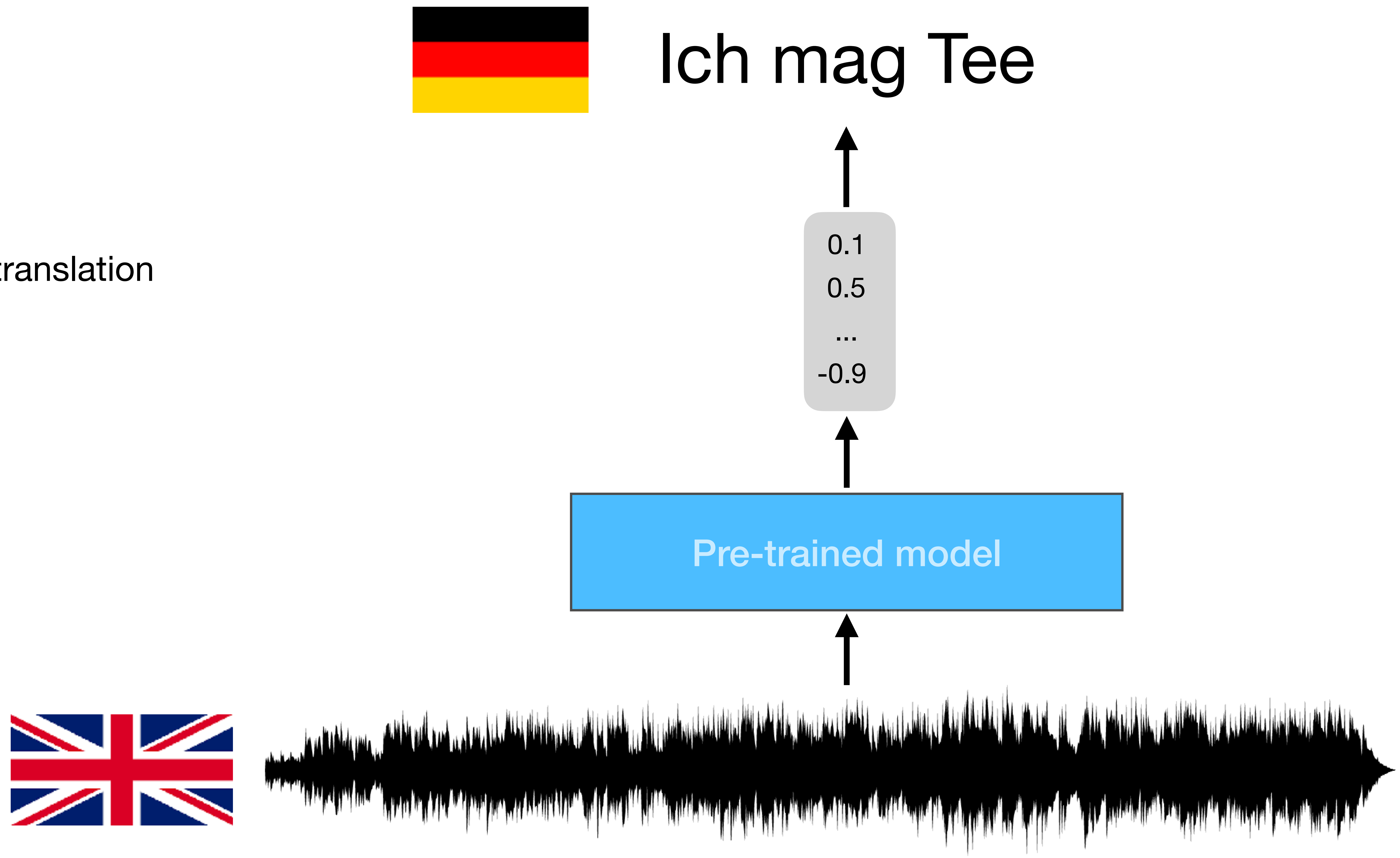
Learning good representations of audio data from unlabeled audio


0.1
0.5
...
-0.9
Pre-trained model

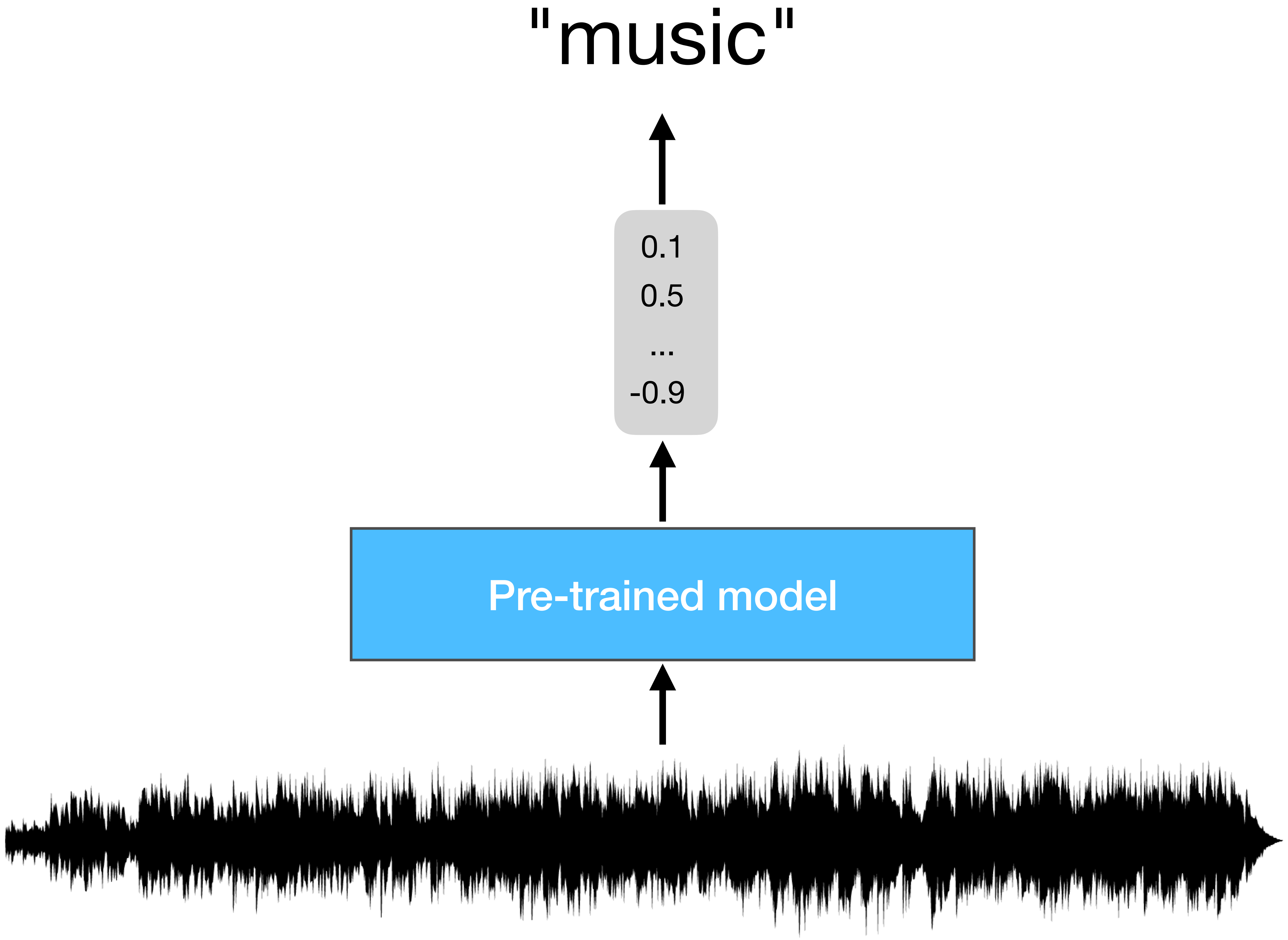
Speech recognition



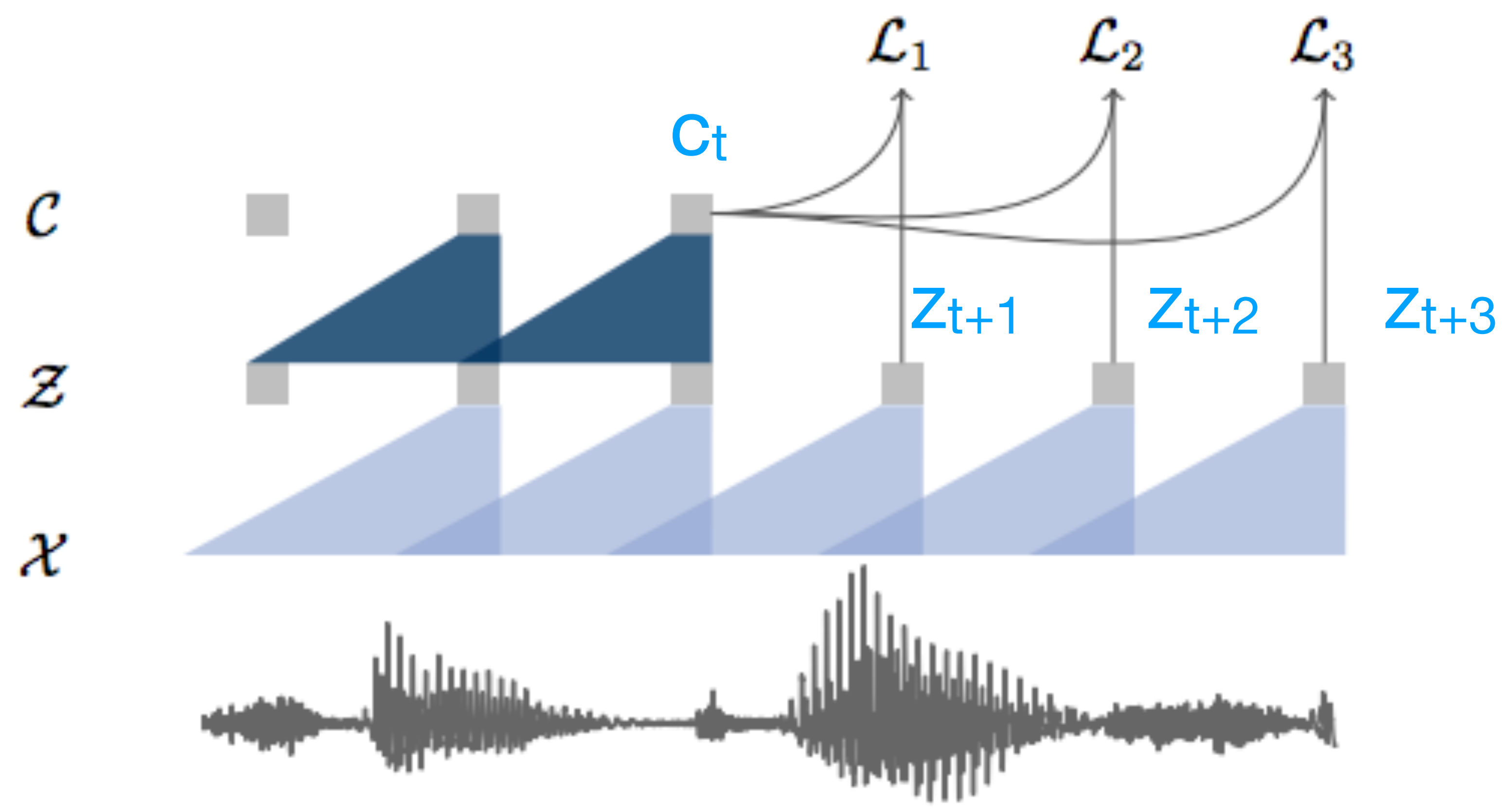
Speech translation



Audio event detection

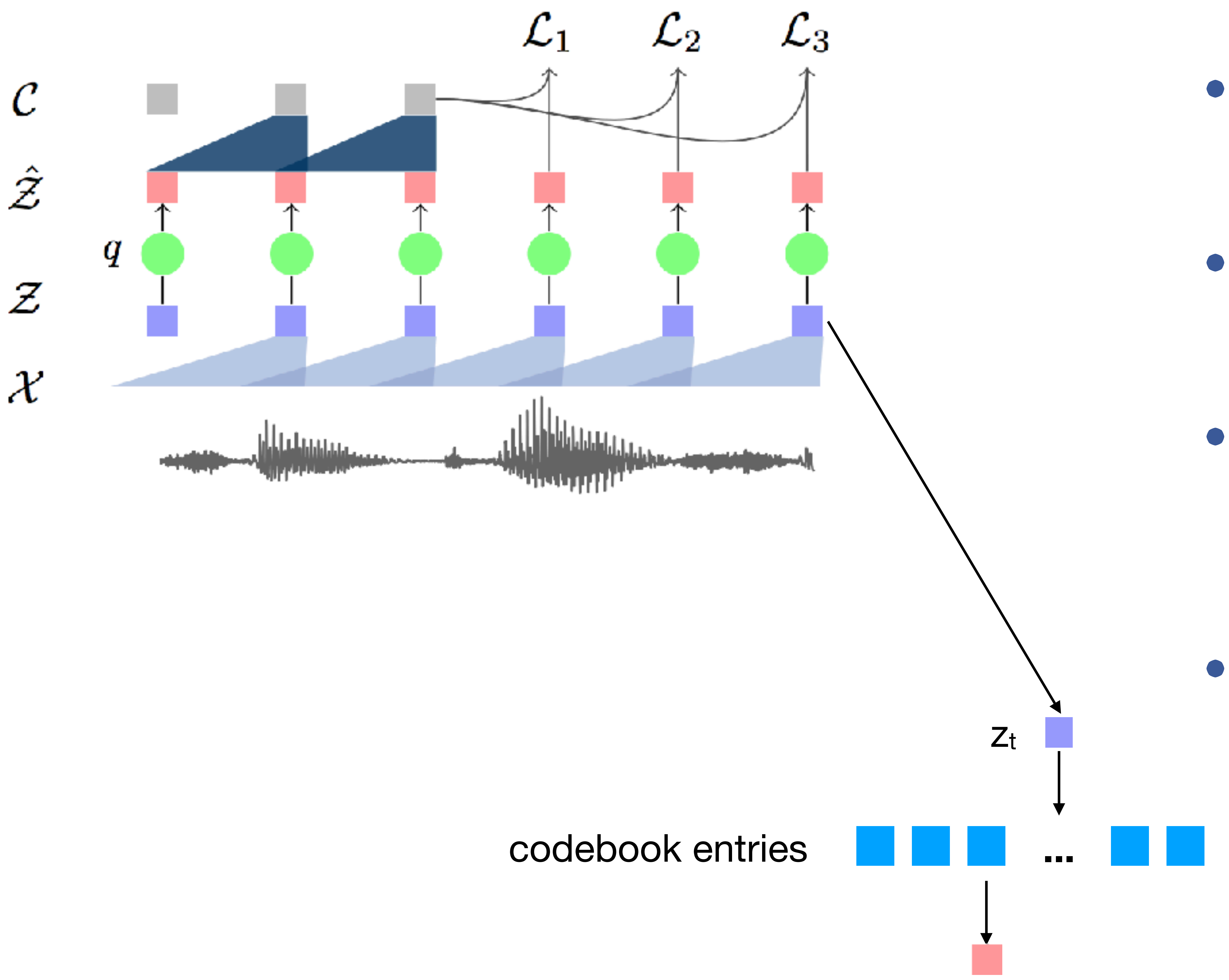


wav2vec: Latent speech audio representations



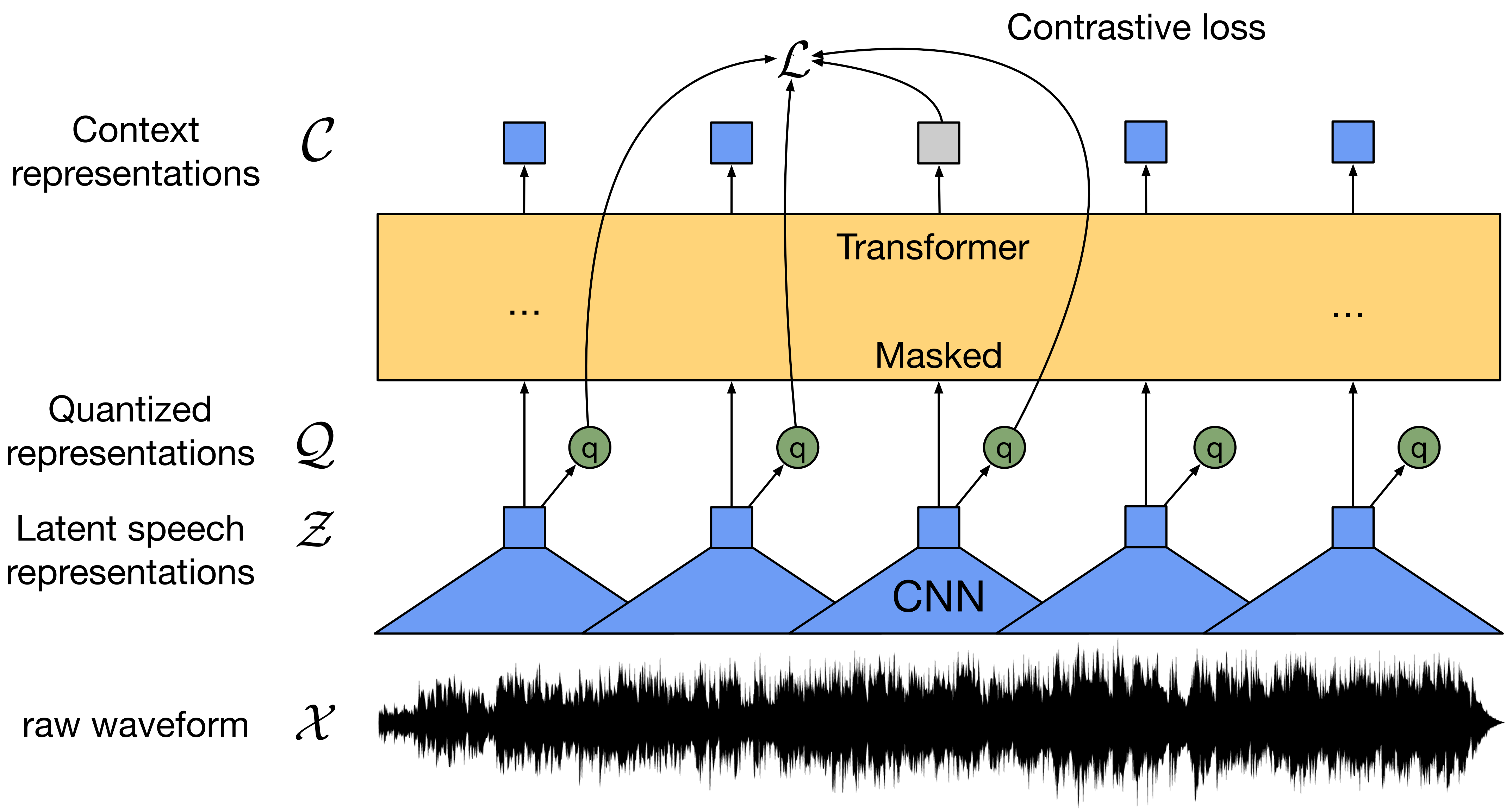
- CNN encodes waveform as latent representations z_t spanning 25ms each
- Another CNN builds context representations c_t of ~300ms
- Training: predict future latents $p(z_{t+1}|c_t)$, $p(z_{t+2}|c_t)$, ...
- Inference: feed c_t into traditional ASR system - instead of logmel etc.

vq-wav2vec: Learning **discrete** latent speech representations



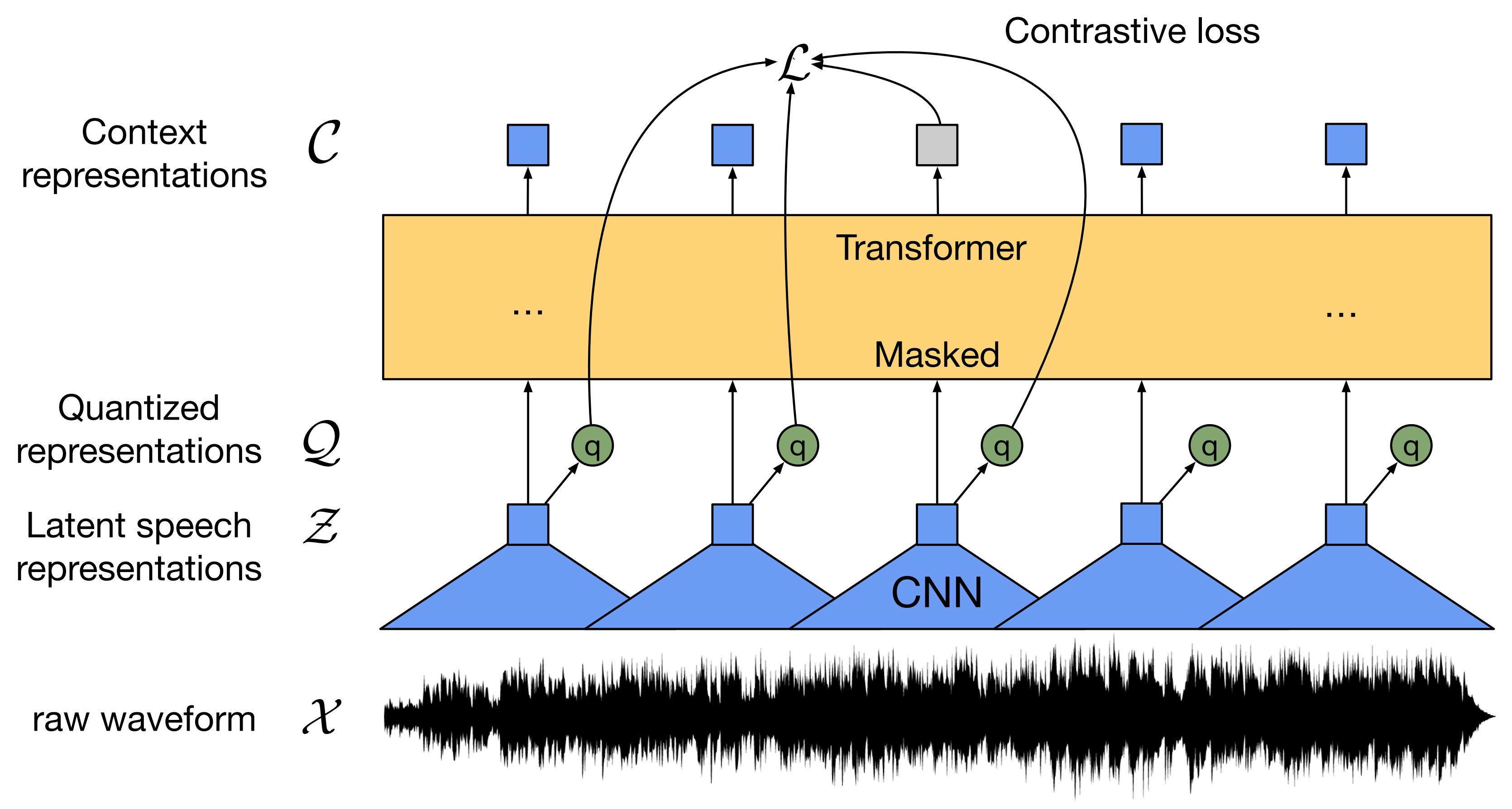
- Human language has a relatively fixed number of possible sounds.
- Mimic this by constraining the latents to a fixed number
- **Vector quantize** the latents = assign each z_t to an entry in a fixed size codebook q by, e.g., online k-means
- Learn an inventory of acoustic units, basic sounds

wav2vec 2.0



- Bi-directional contextualized representations
- Vector quantized targets for training

Objective



Cosine similarity

Context representation

Discrete latent speech representation

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}}\sim\mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

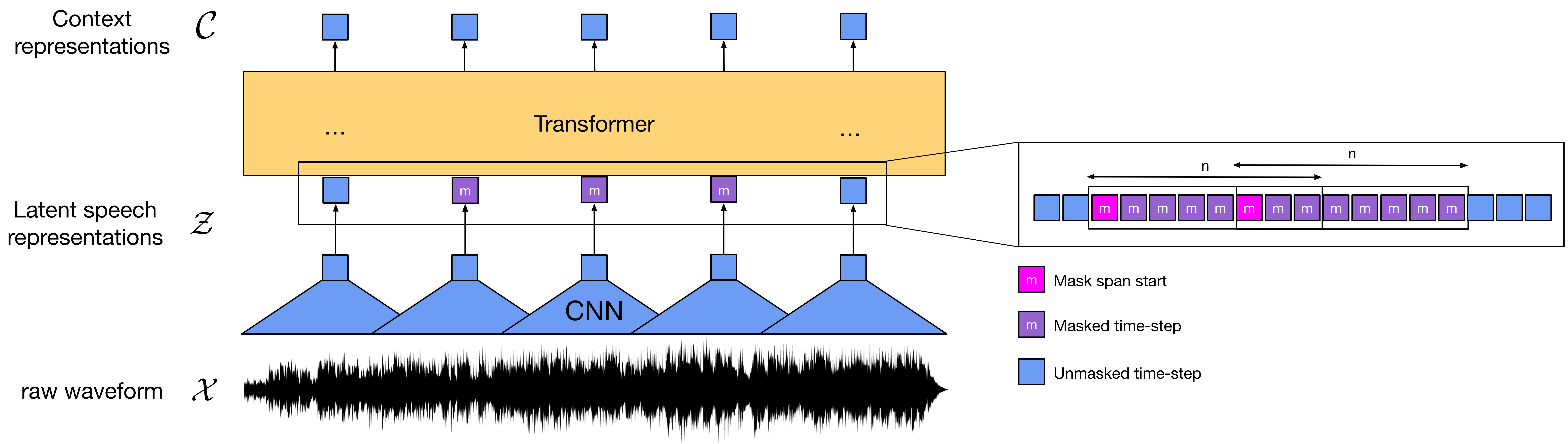
Negative samples

Temperature

Codebook diversity penalty to encourage more codes to be used

Masking

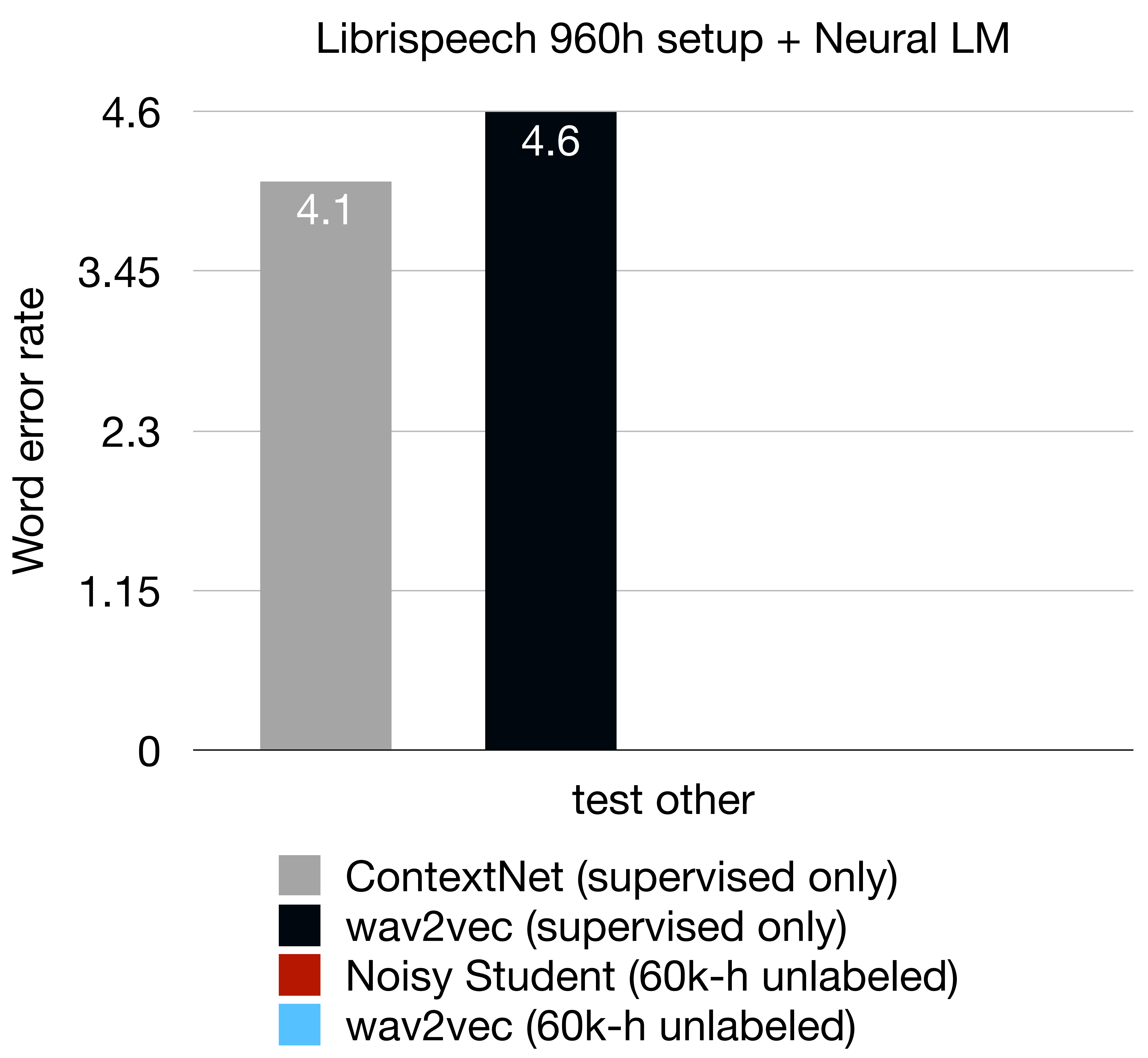
- Sample starting points for masks without replacement, then expand to 10 time-steps (1 time-step is 25ms but 10ms stride)
- Spans can overlap
- For a 15s sample, ~49% of the time-steps masked with an average span length of ~300ms



Fine-tuning

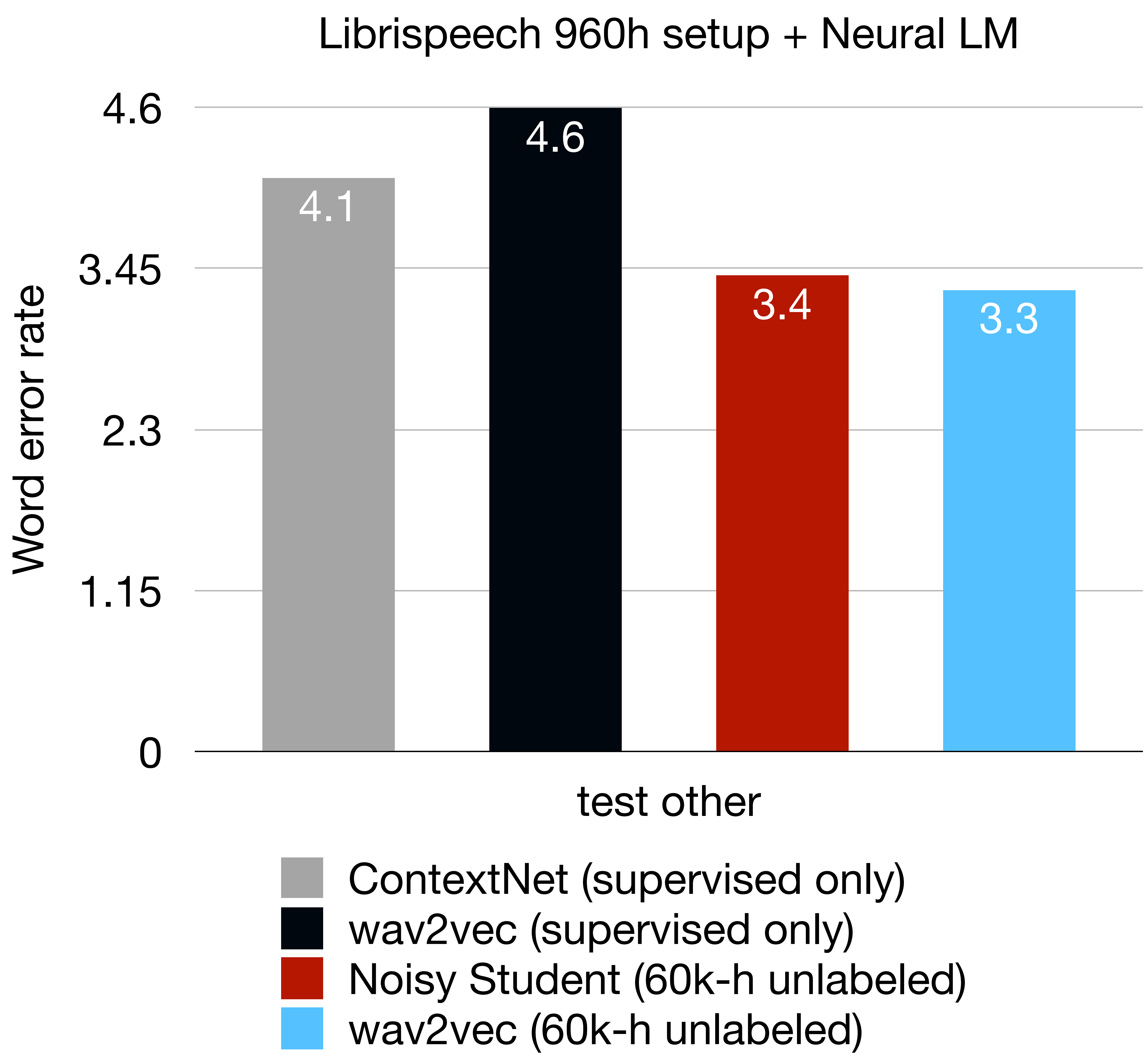
- Add a single linear projection on top into target vocab and train with CTC loss with a low learning rate (CNN encoder is not trained).
- Use modified SpecAugment in latent space to prevent early overfitting
- Uses wav2letter decoder with the official 4gram LM and Transformer LM

Results



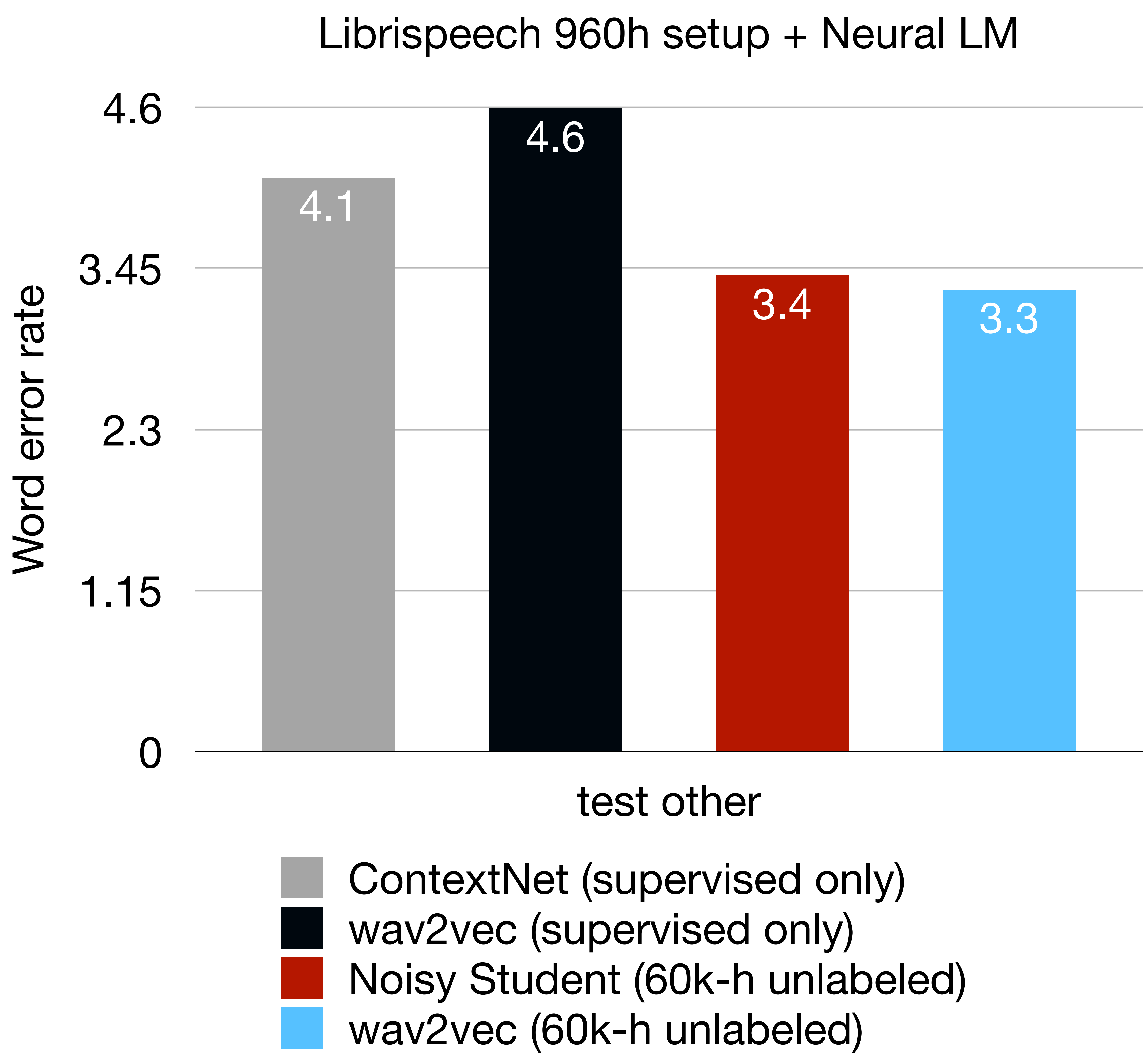
Librispeech 960h setup + Neural LM
4.6
3.45
2.3
1.15
0
Word error rate
4.1
4.6
test other
ContextNet (supervised only)
wav2vec (supervised only)
Noisy Student (60k-h unlabeled)
wav2vec (60k-h unlabeled)

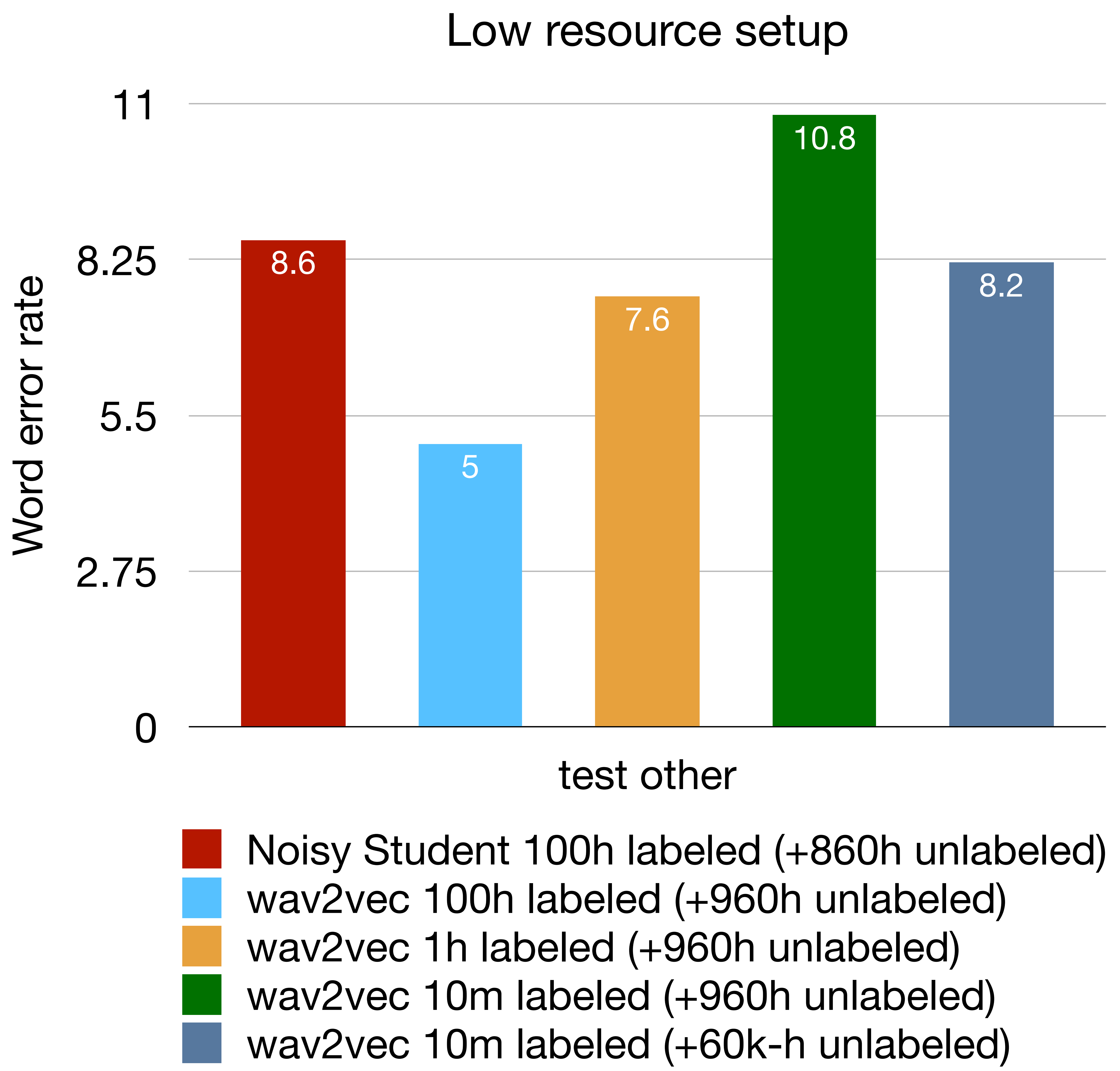
Results



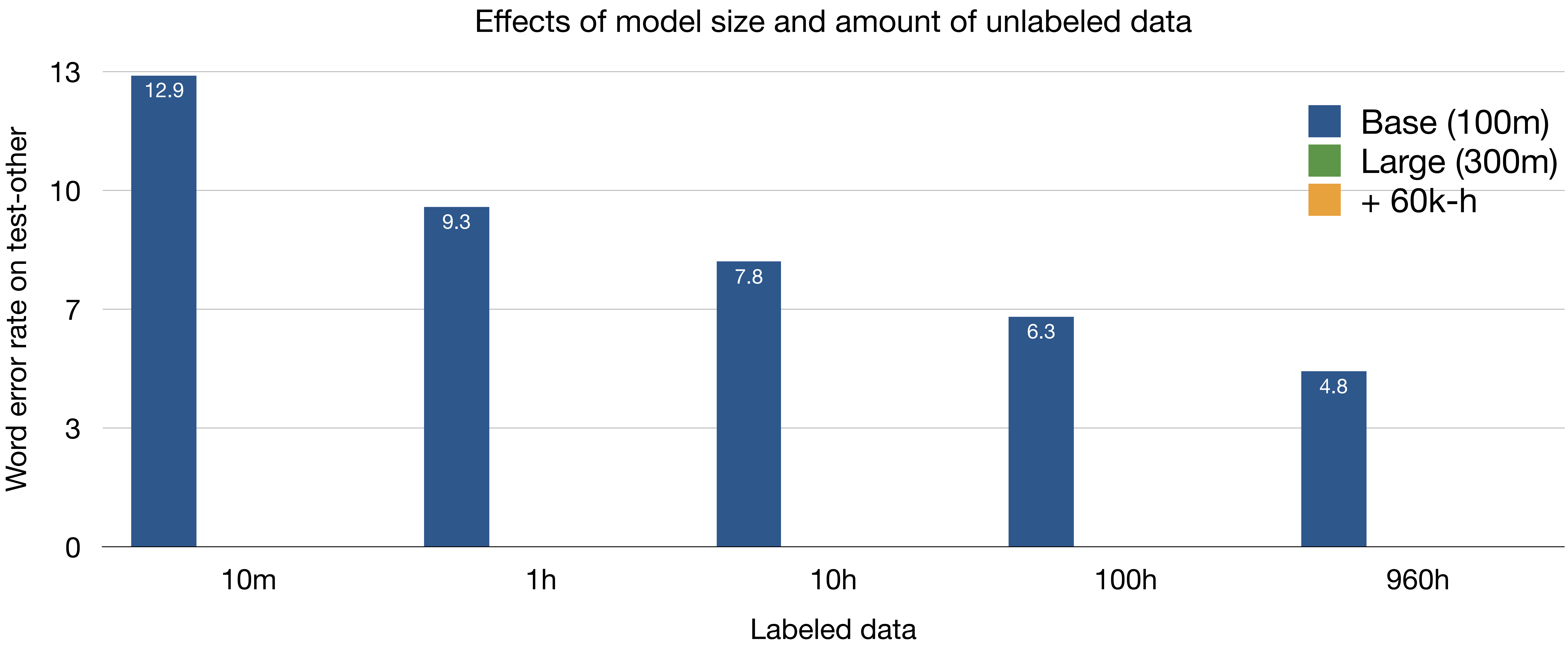
Librispeech 960h setup + Neural LM
Word error rate
4.6
3.45
2.3
1.15
0
4.1
4.6
3.4
3.3
test other
ContextNet (supervised only)
wav2vec (supervised only)
Noisy Student (60k-h unlabeled)
wav2vec (60k-h unlabeled)

Results

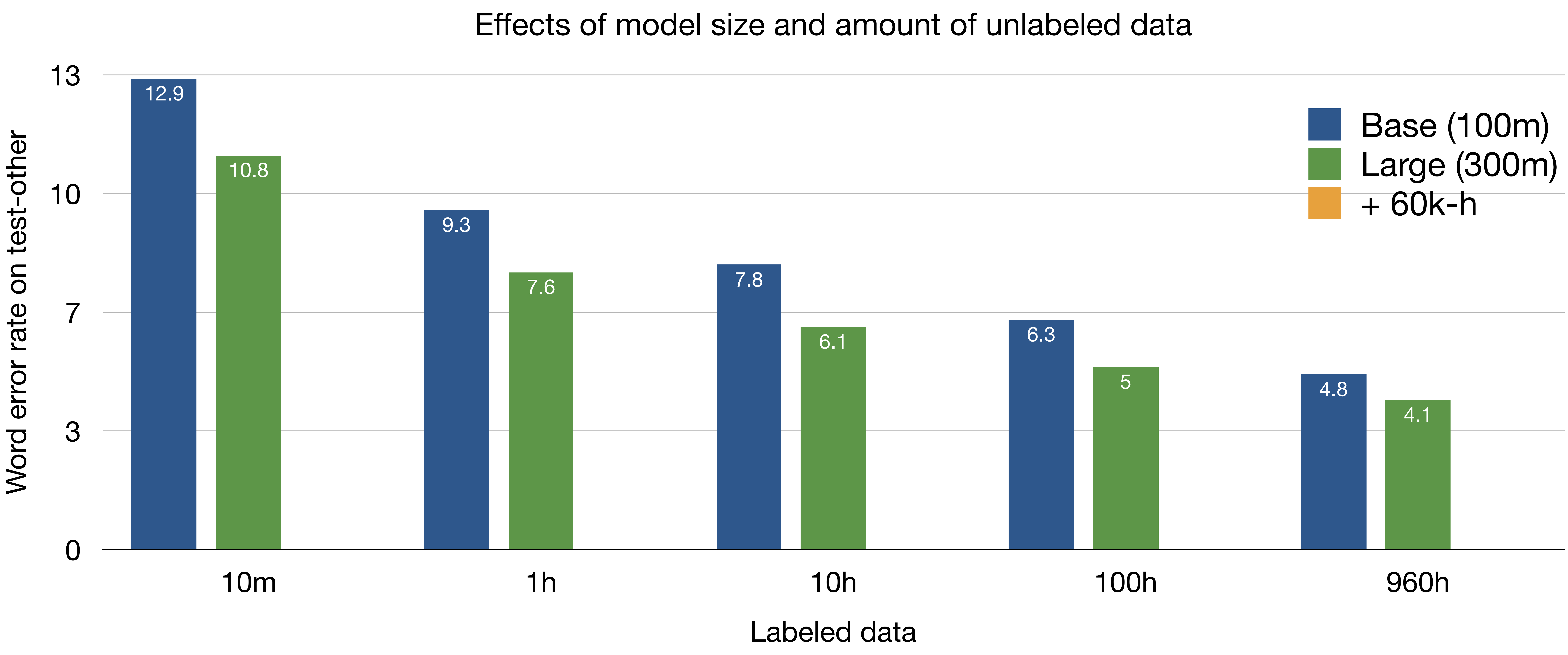

Librispeech 960h setup + Neural LM
Word error rate
4.6
3.45
2.3
1.15
0
4.1
4.6
3.4
3.3
test other
ContextNet (supervised only)
wav2vec (supervised only)
Noisy Student (60k-h unlabeled)
wav2vec (60k-h unlabeled)


Low resource setup
Word error rate
11
8.25
5.5
2.75
0
8.6
5
7.6
10.8
8.2
test other
Noisy Student 100h labeled (+860h unlabeled)
wav2vec 100h labeled (+960h unlabeled)
wav2vec 1h labeled (+960h unlabeled)
wav2vec 10m labeled (+960h unlabeled)
wav2vec 10m labeled (+60k-h unlabeled)

Results

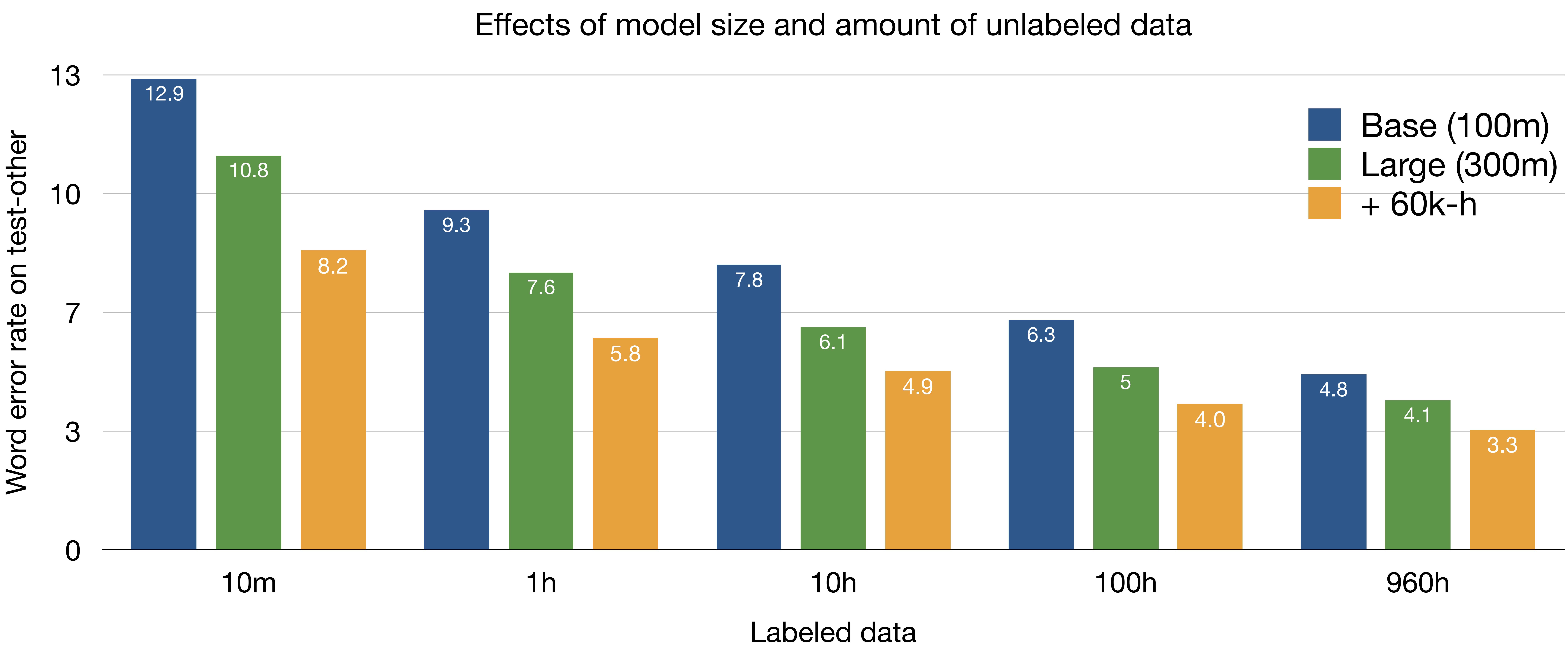

Effects of model size and amount of unlabeled data
Word error rate on test-other
13
10
7
3
0
12.9
9.3
7.8
6.3
4.8
10m
1h
10h
100h
960h
Labeled data
Base (100m)
Large (300m)
+ 60k-h

Results



Effects of model size and amount of unlabeled data
Word error rate on test-other
13
10
7
3
0
12.9
10.8
9.3
7.6
7.8
6.1
6.3
5
4.8
4.1
Base (100m)
Large (300m)
+ 60k-h
10m
1h
10h
100h
960h
Labeled data

Results



Effects of model size and amount of unlabeled data
Word error rate on test-other
13
10
7
3
0
12.9
10.8
8.2
9.3
7.6
5.8
7.8
6.1
4.9
6.3
5
4.0
4.8
4.1
3.3
10m
1h
10h
100h
960h
Labeled data
Base (100m)
Large (300m)
+ 60k-h

Examples (10 min labeled data)

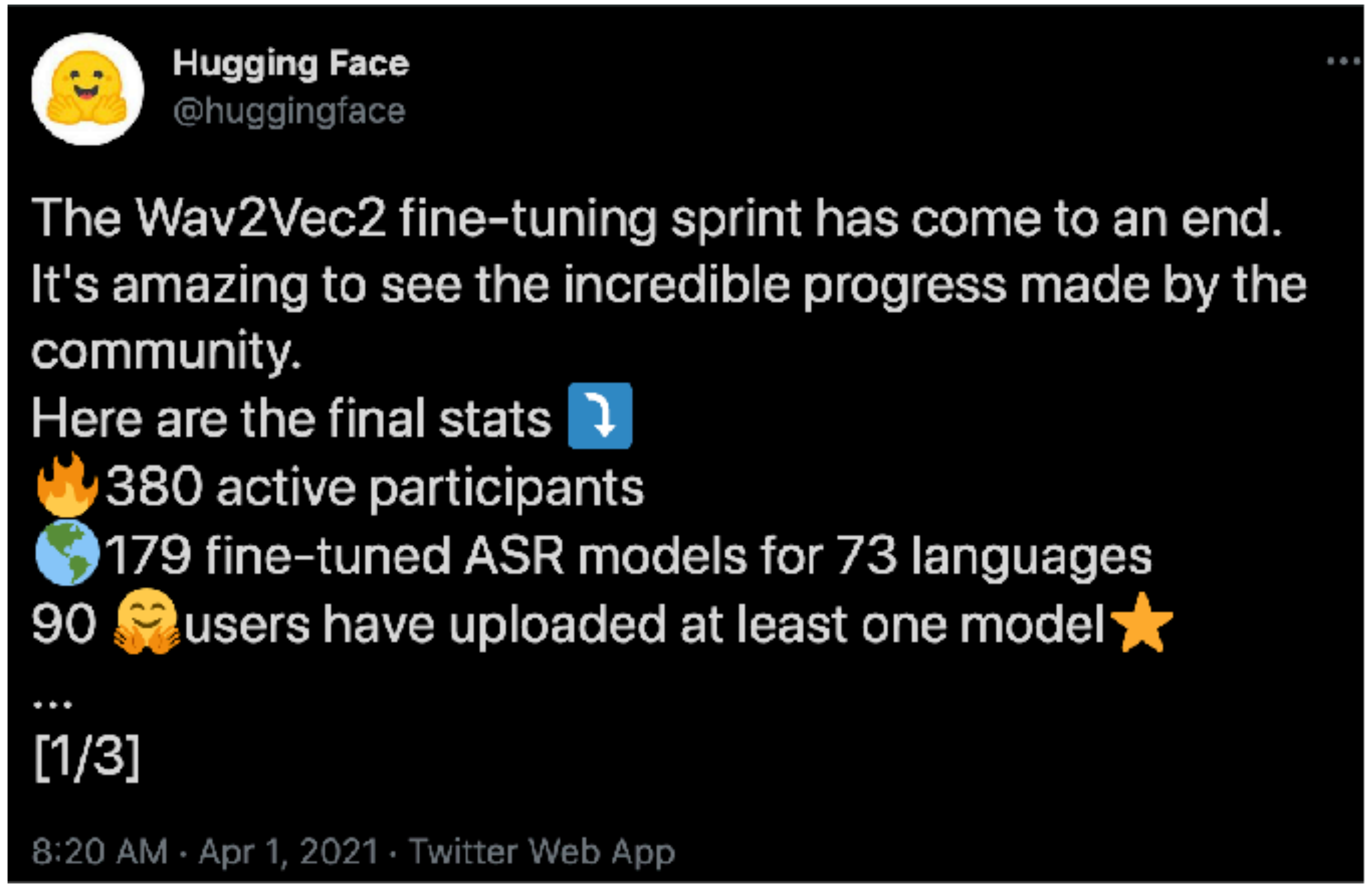
HYP (no LM): she SESED and LUCHMAN GAIVE A SENT won by her GENTAL argument
HYP (w/ LM): she ceased and LUCAN gave assent won by her gentle argument
REF: she ceased and lakshman gave assent won by her gentle argument

HYP (no LM): but NOT WITH STANDING this boris EMBRAED him in a QUIAT FRENDLY way and CISED him THRE times
HYP (w/ LM): but NOT WITHSTANDING this boris embraced him in a quiet friendly way and kissed him three times
REF: but notwithstanding this boris embraced him in a quiet friendly way and kissed him three times

wav2vec on HuggingFace

- HuggingFace is a popular NLP model zoo

- HuggingFace community fine-tuned our models to do speech recognition in 73 languages.



Pre-training and self-training

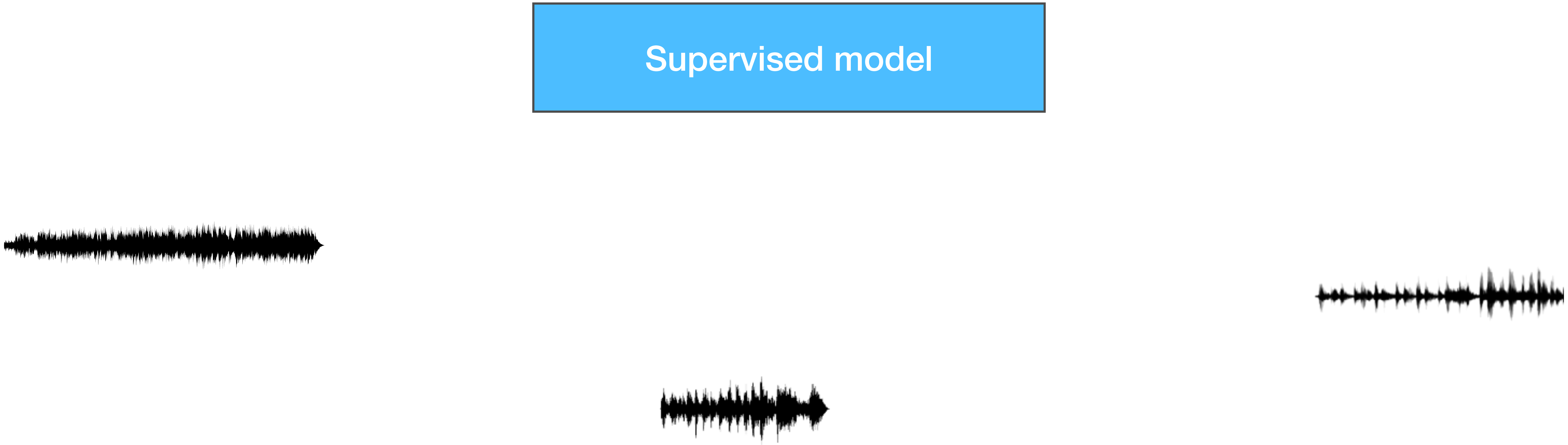
Pre-training and self-training

- Self-training very successful in speech recognition: generate pseudo-labels

Supervised model

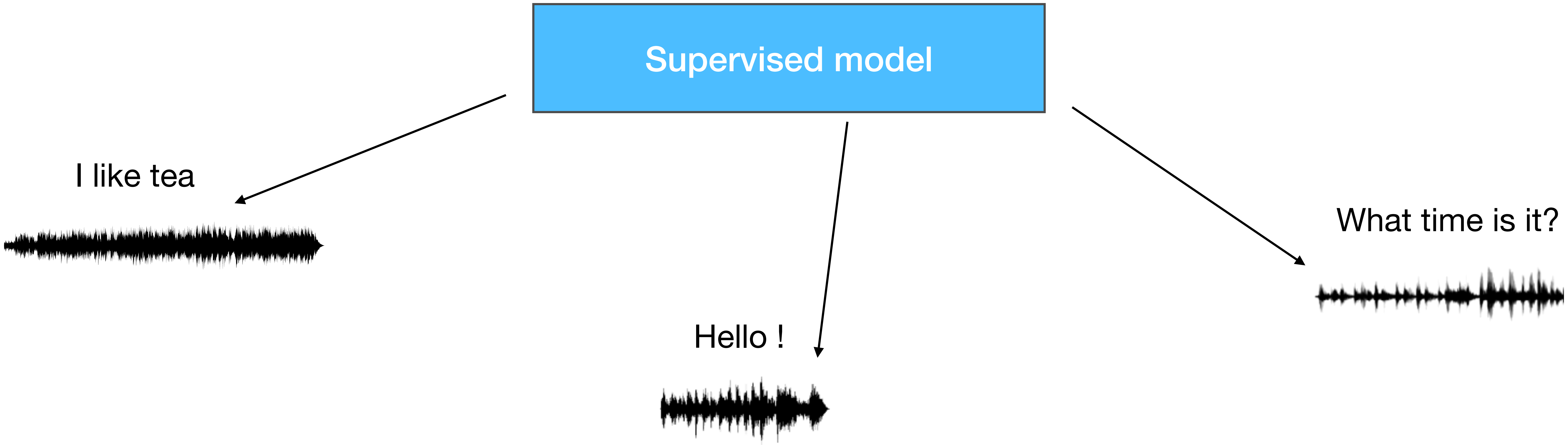
Pre-training and self-training

- Self-training very successful in speech recognition: generate pseudo-labels



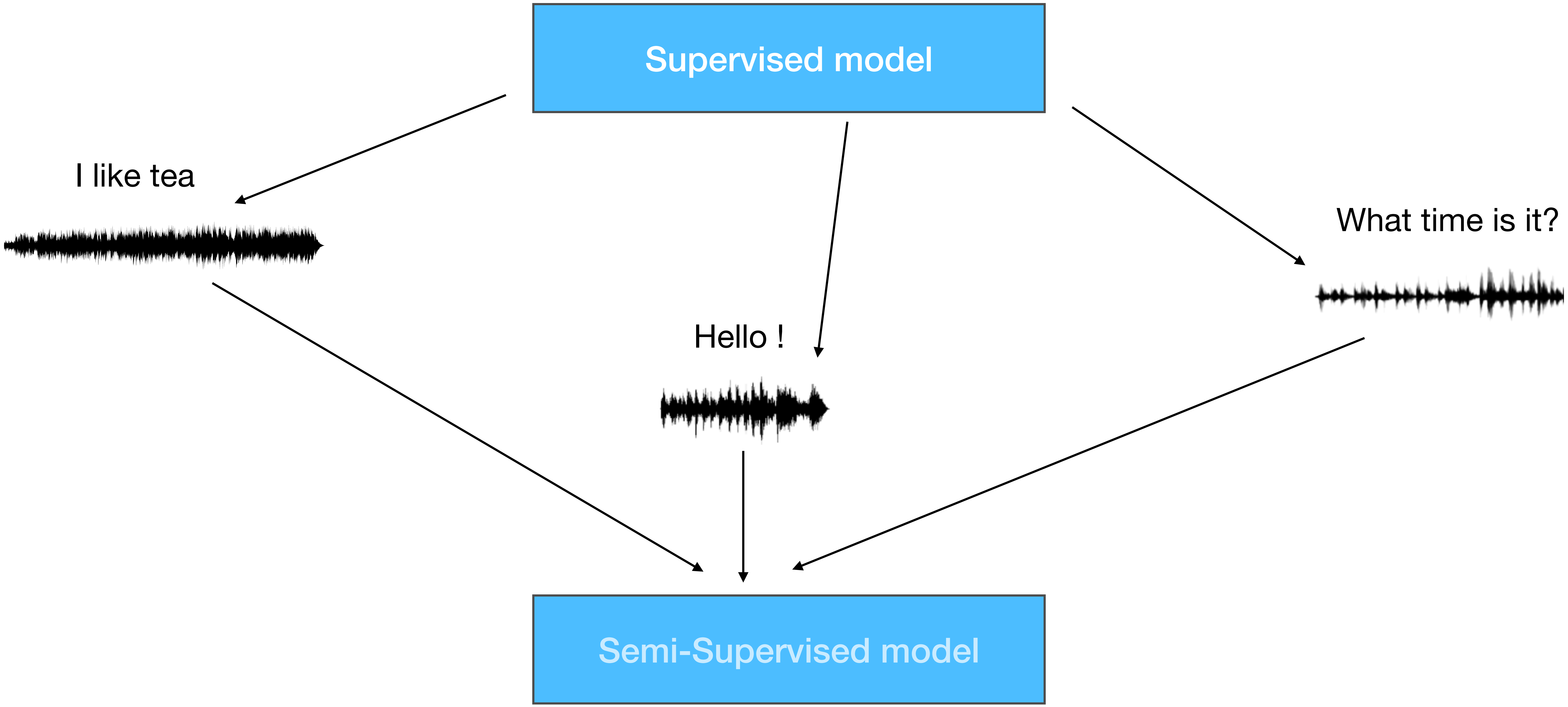
Pre-training and self-training

- Self-training very successful in speech recognition: generate pseudo-labels



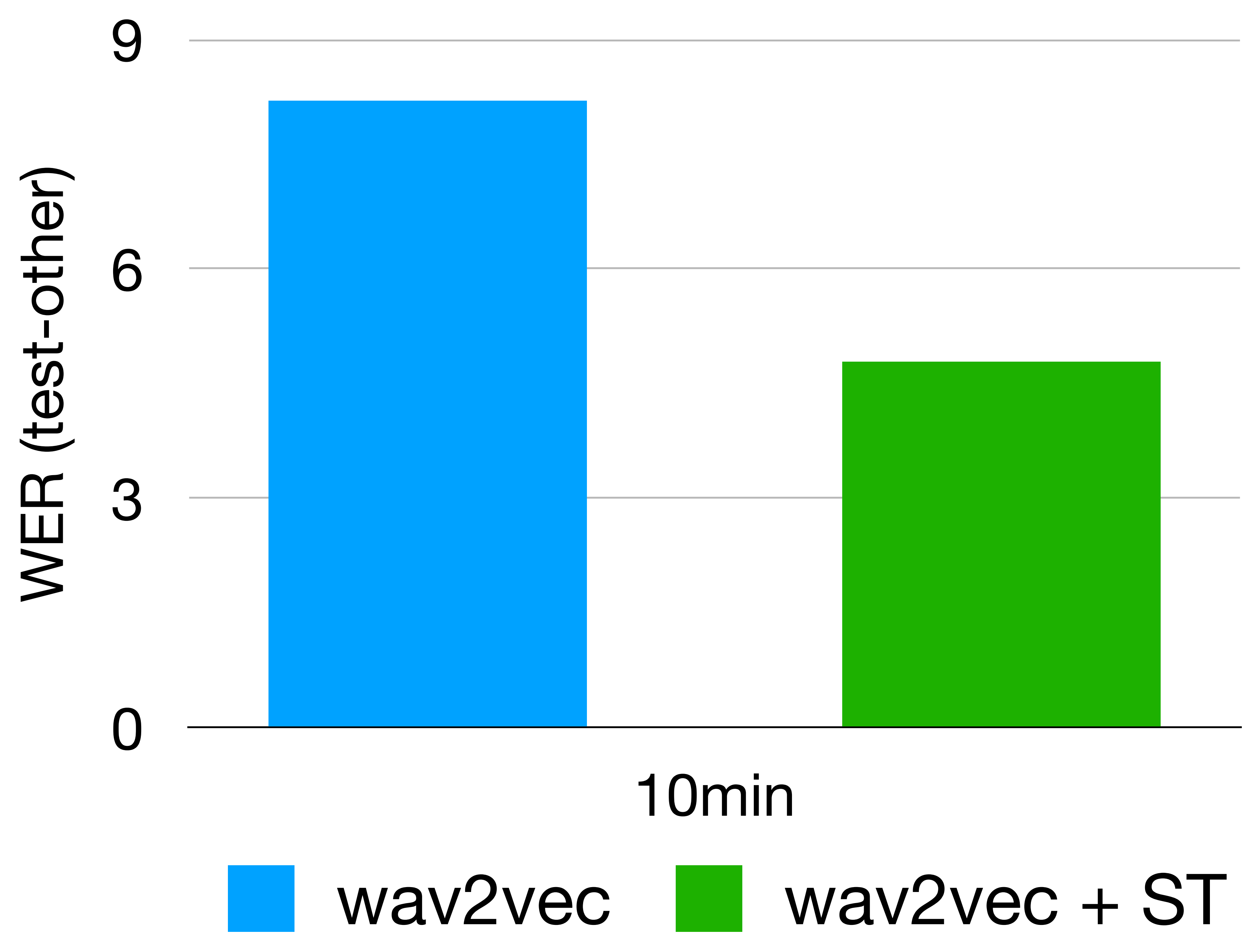
Pre-training and self-training

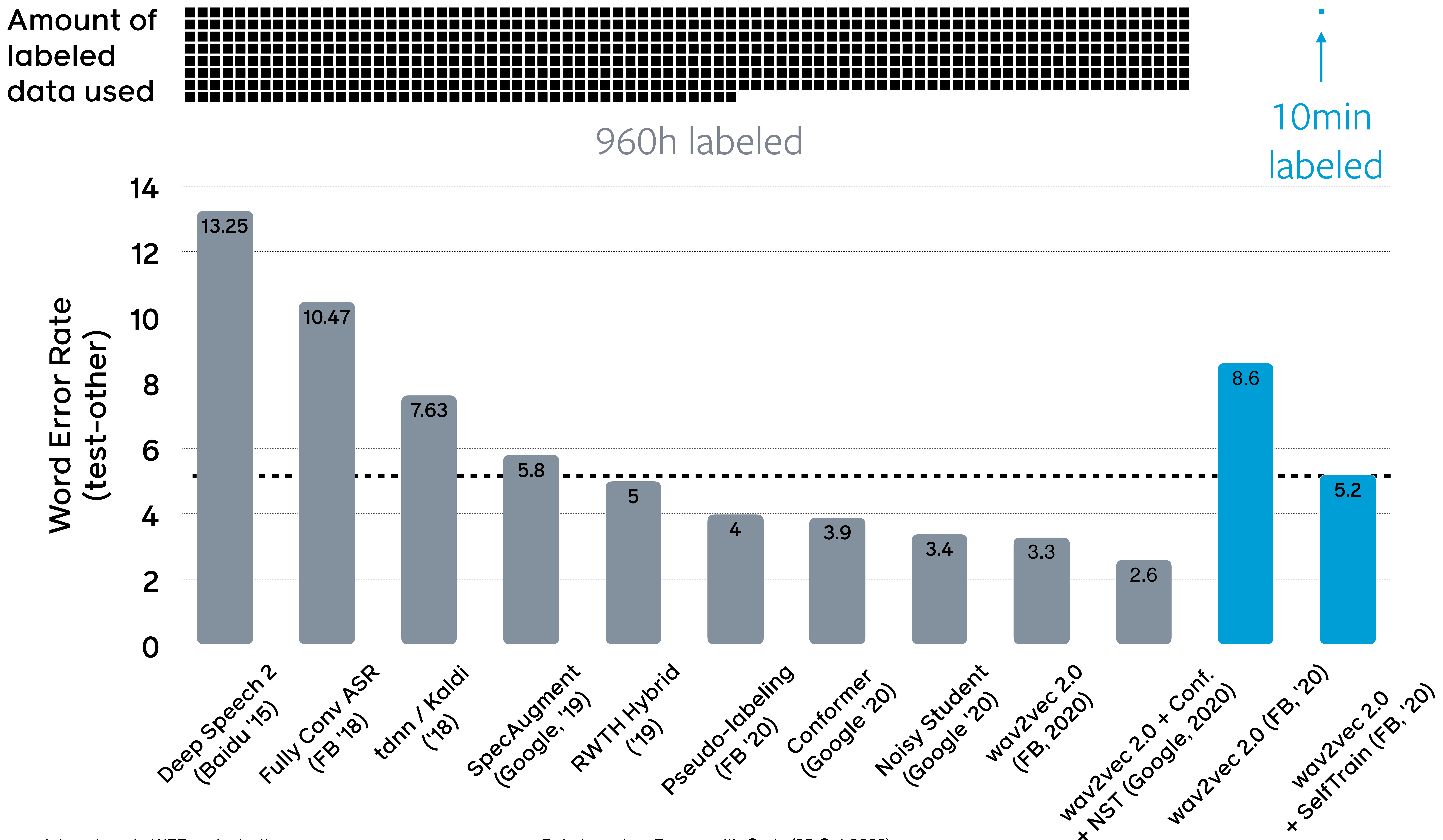
- Self-training very successful in speech recognition: generate pseudo-labels



Pre-training and self-training

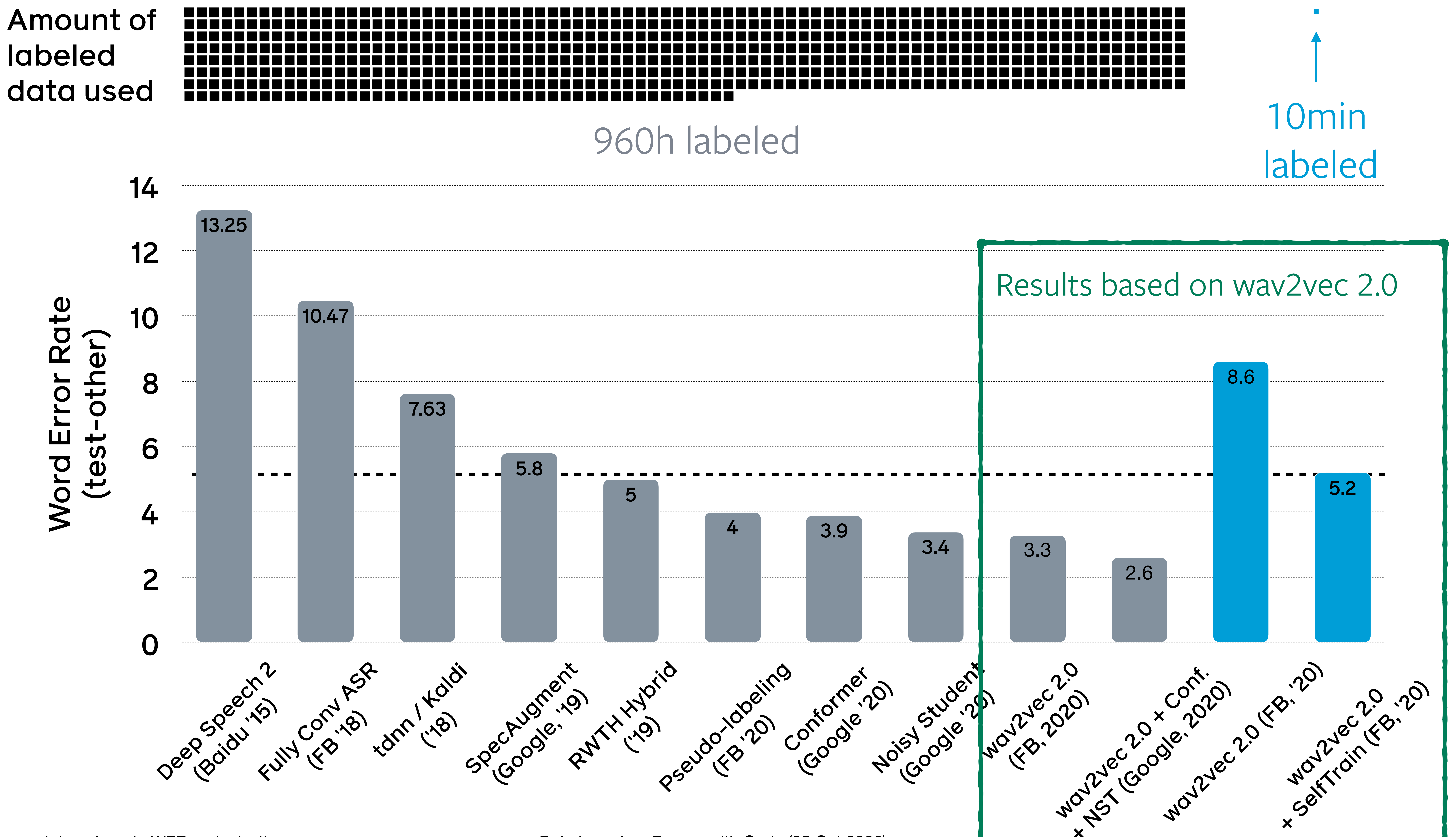
- Self-training very successful in speech recognition: generate pseudo-labels
- Do both have the same effect?
- Recipe: pre-train on the unlabeled data, pseudo-label, fine-tune pre-trained model





Librispeech benchmark, WER on test-other

Data based on Papers with Code (25 Oct 2020)



Librispeech benchmark, WER on test-other

Data based on Papers with Code (25 Oct 2020)

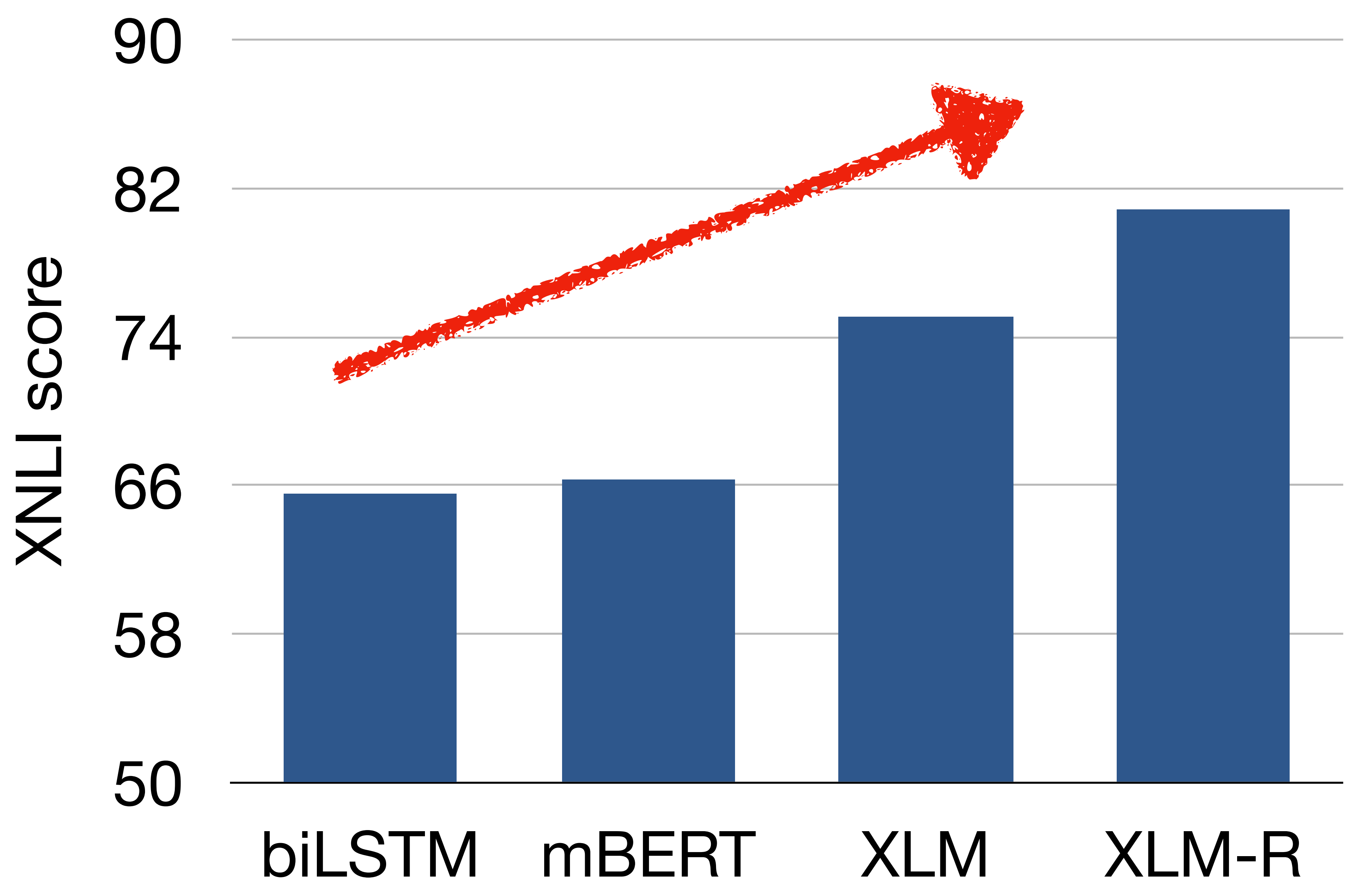
XLSR: cross lingual speech representation learning with wav2vec

Why *cross-lingual* self-supervised learning

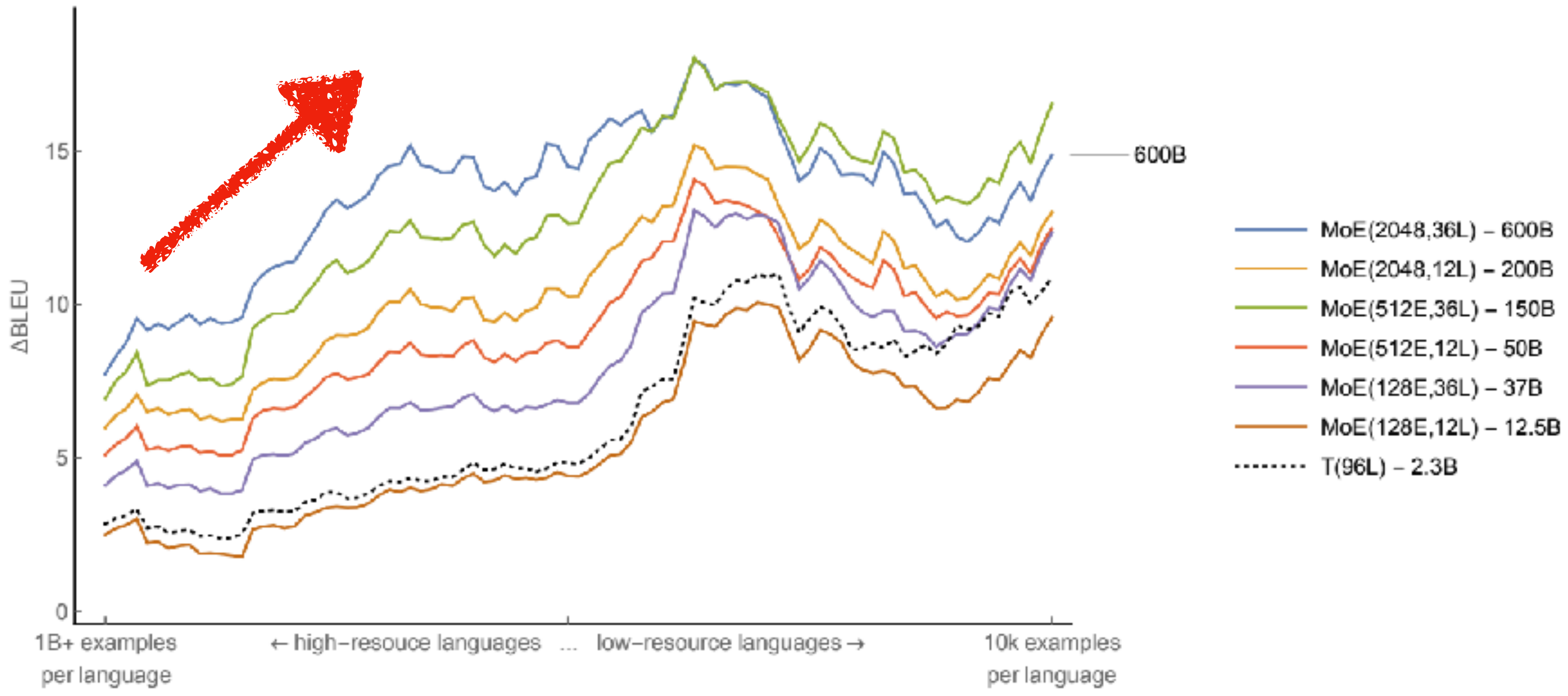
- Little labeled data -> little unlabeled data
- Leverage unlabeled data from high-resource languages
- To improve performance on low-resource languages
- One model for each of the 6500 languages, for each domain? No.
- Instead: one pertained model for all languages

Meanwhile in multilingual research

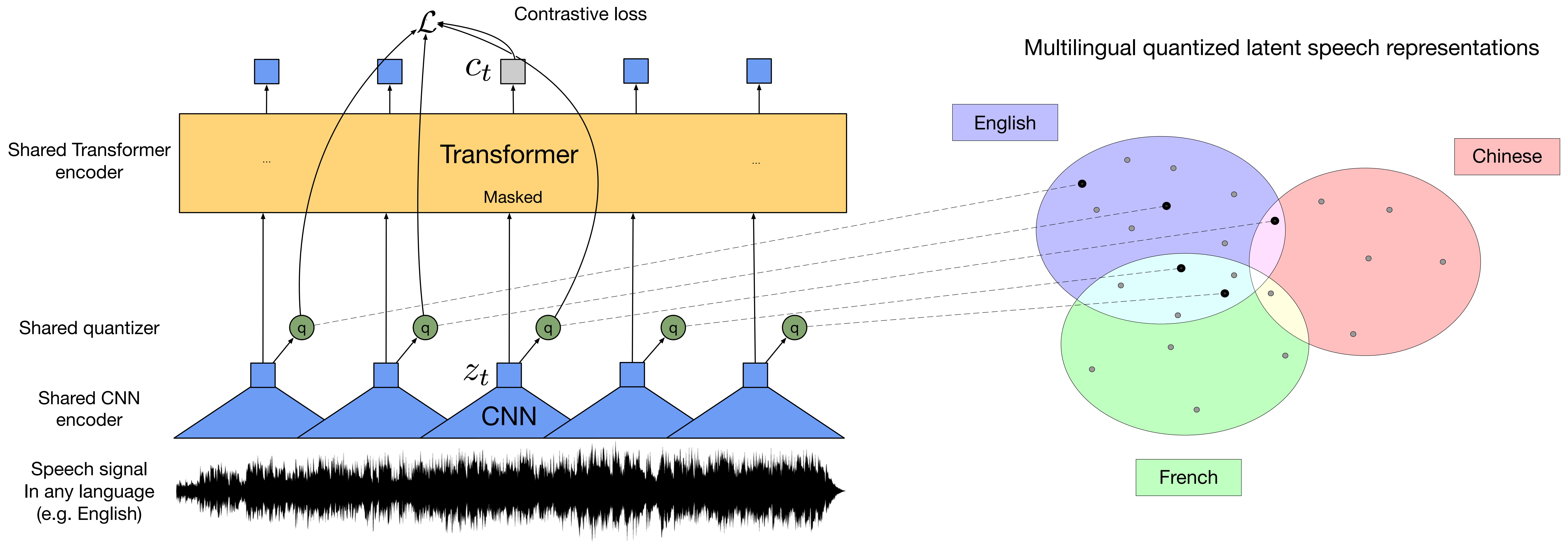
Cross-lingual understanding (XLU)



Multilingual machine translation



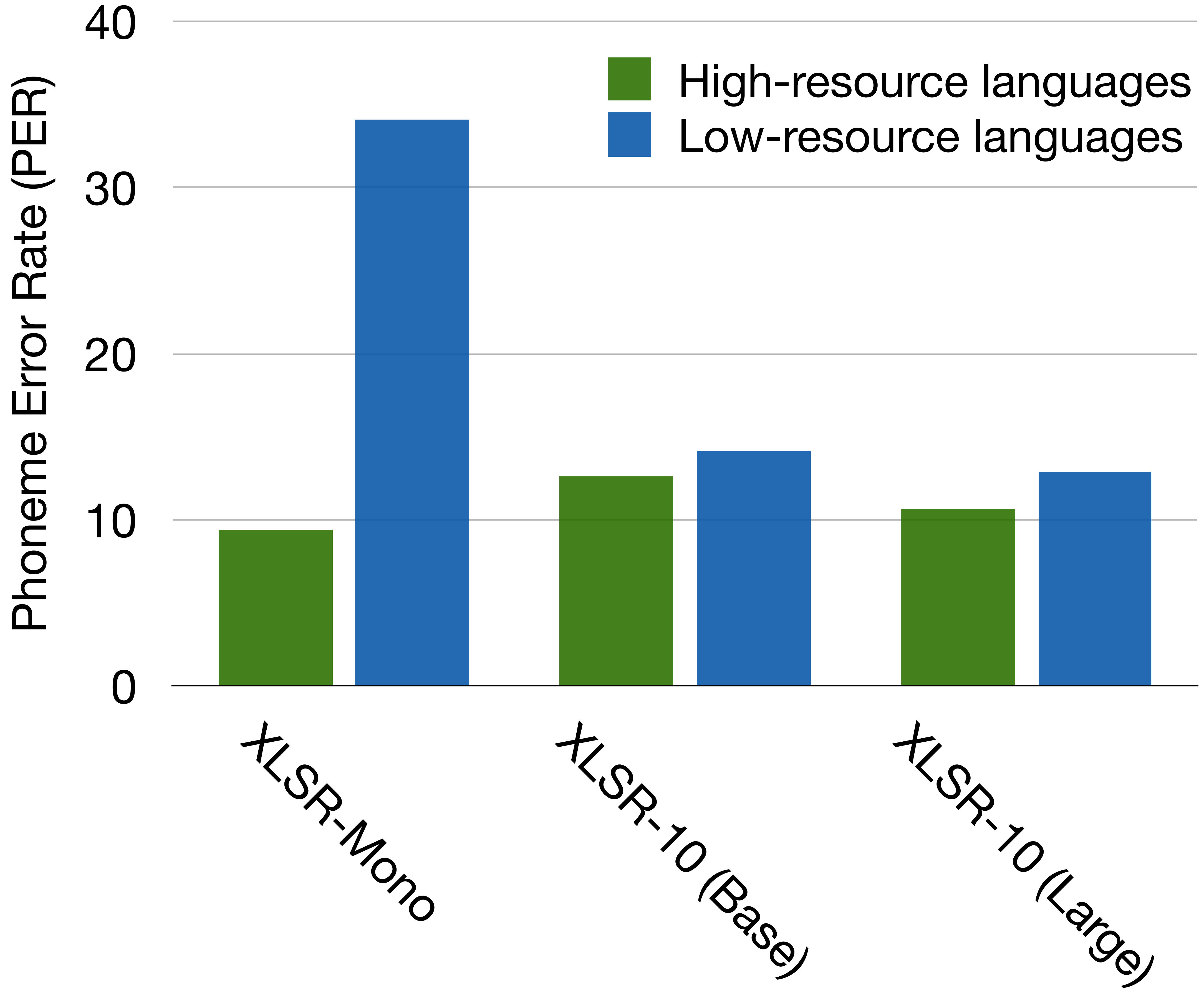
XLSR: cross lingual speech representation learning with wav2vec



XLSR: Results - cross-lingual transfer

XLSR significantly outperforms previously published approaches on CommonVoice/BABEL

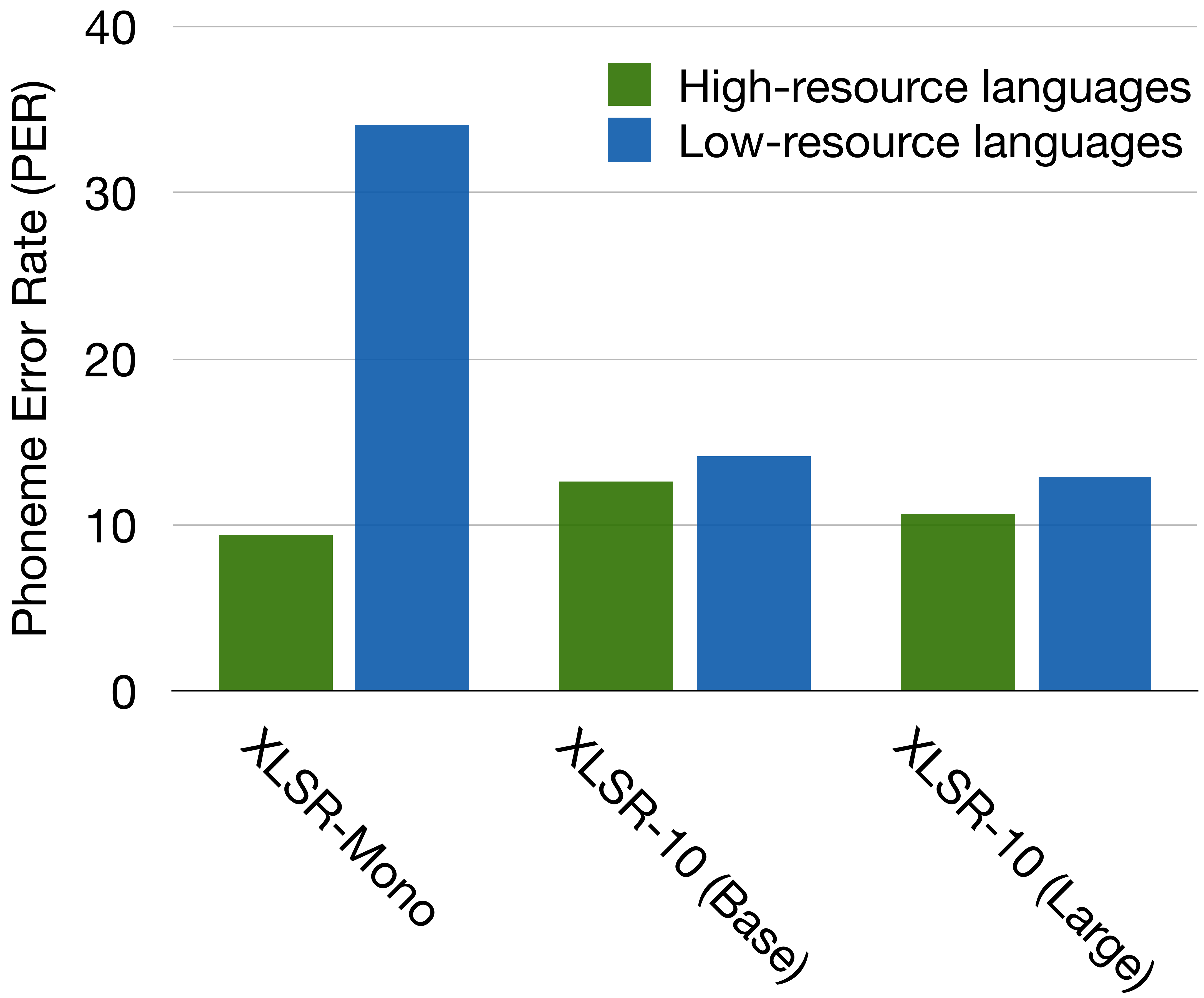
CommonVoice results:



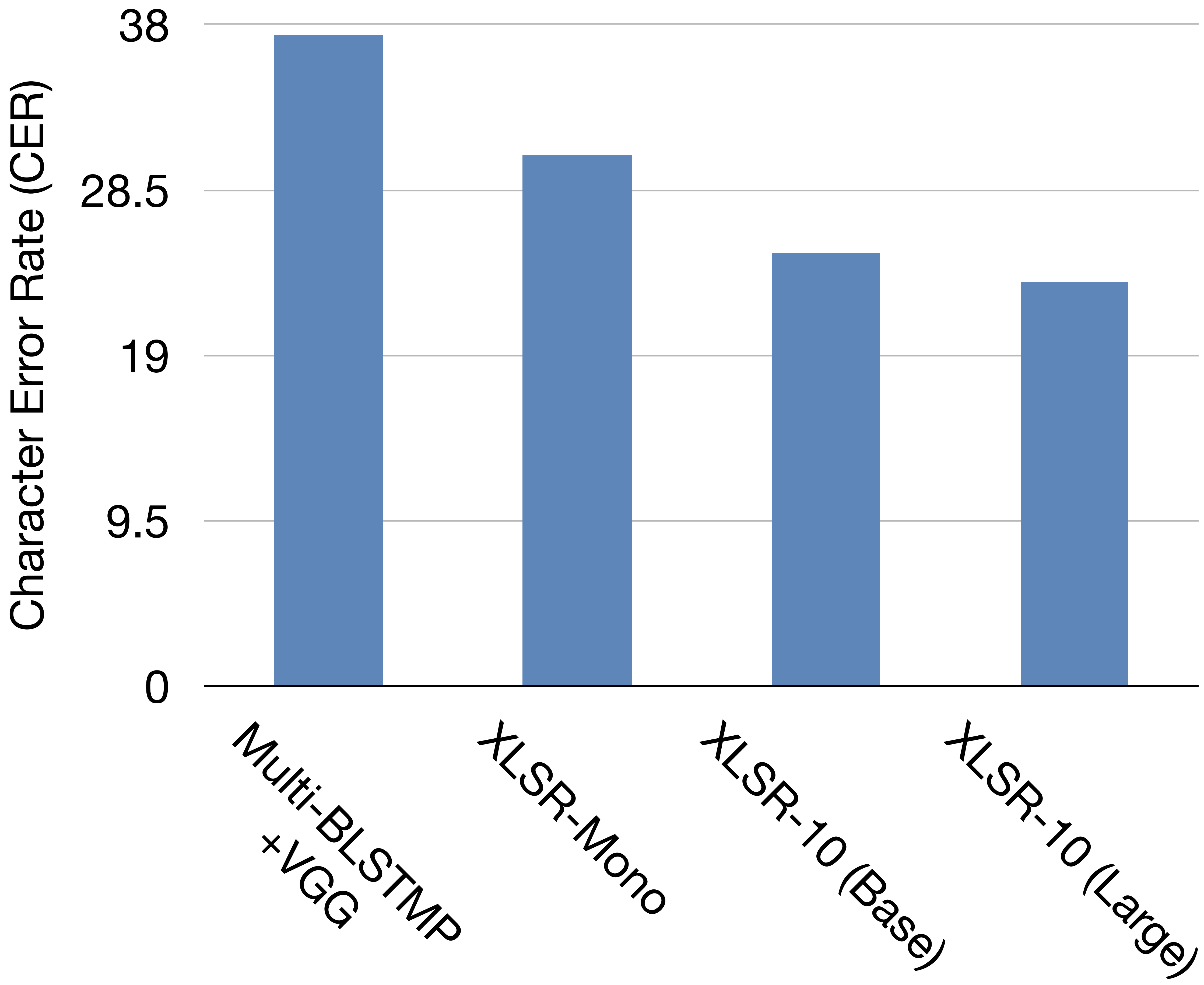
XLSR: Results - cross-lingual transfer

XLSR significantly outperforms previously published approaches on CommonVoice/BABEL

CommonVoice results:



BABEL (average) results:



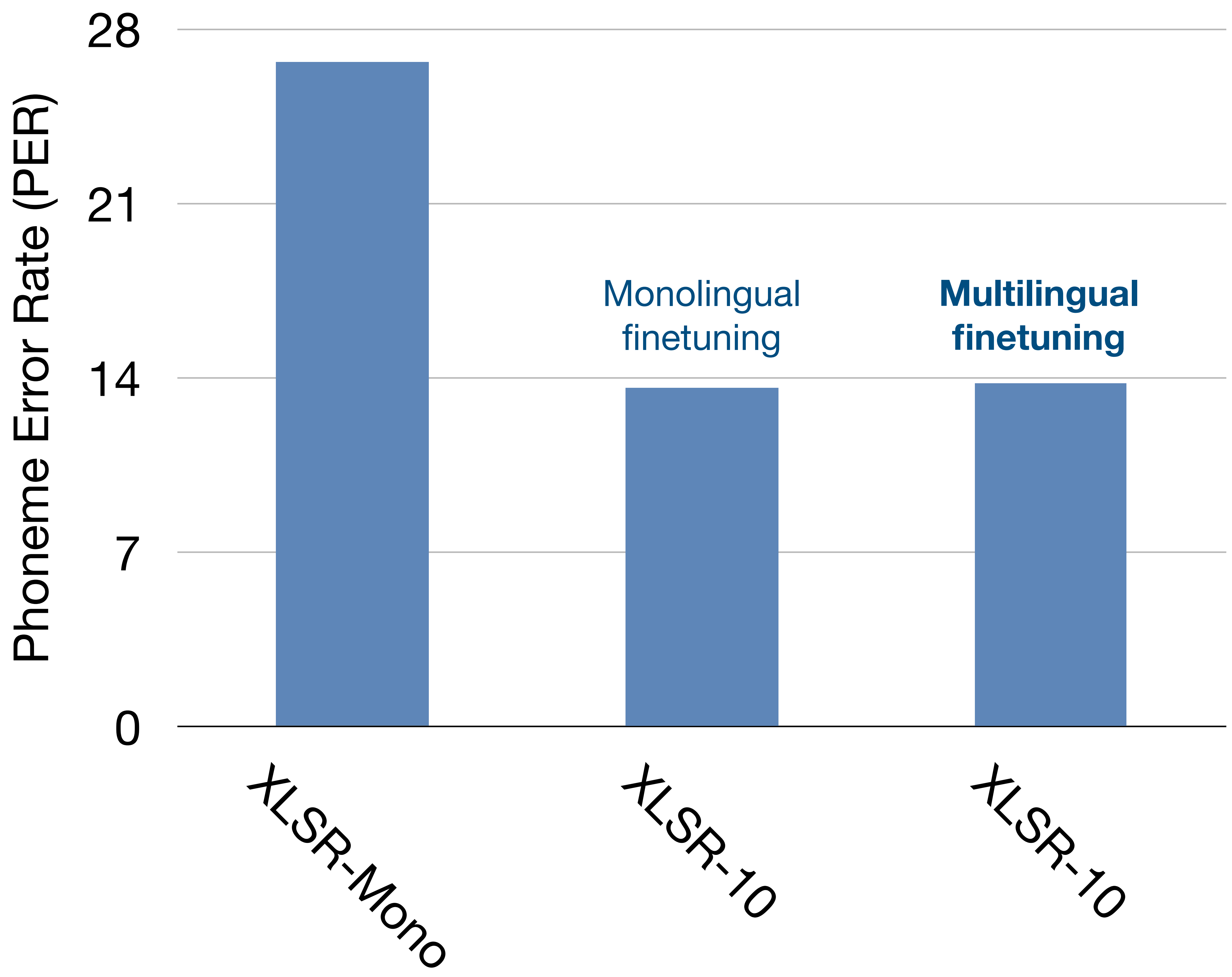
XLSR: Results - multilingual fine-tuning

Multilingual finetuning leads to *one model for all languages* with little loss in performance

XLSR: Results - multilingual fine-tuning

Multilingual finetuning leads to *one model for all languages* with little loss in performance

CommonVoice results:



XLSR: Results - impact of language similarity

Language similarity plays an important role in cross-lingual transfer

Similar higher-resource language data helps the most for low-resource language

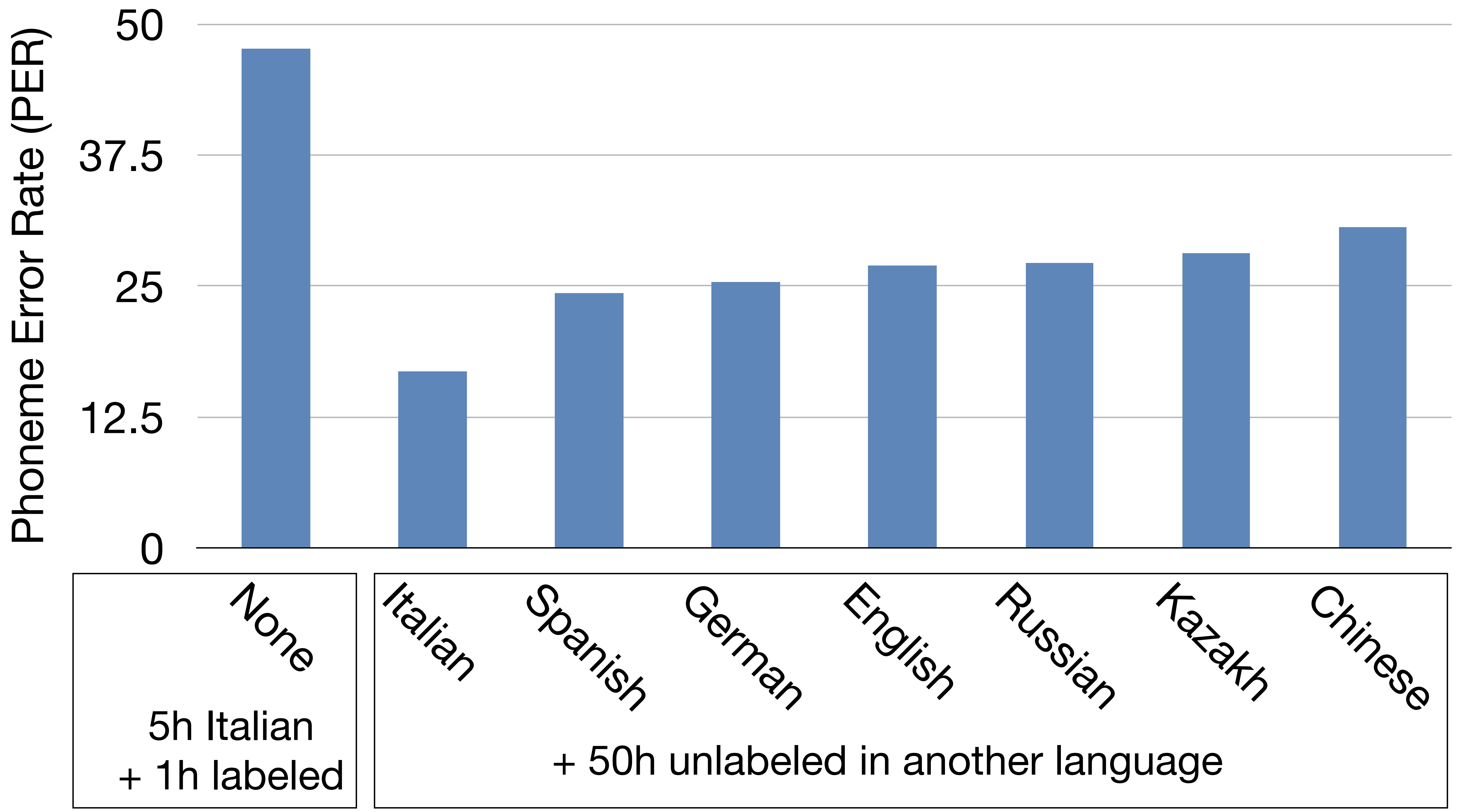
5h Italian
+ 1h labeled

+ 50h unlabeled in another language

XLSR: Results - impact of language similarity

Language similarity plays an important role in cross-lingual transfer

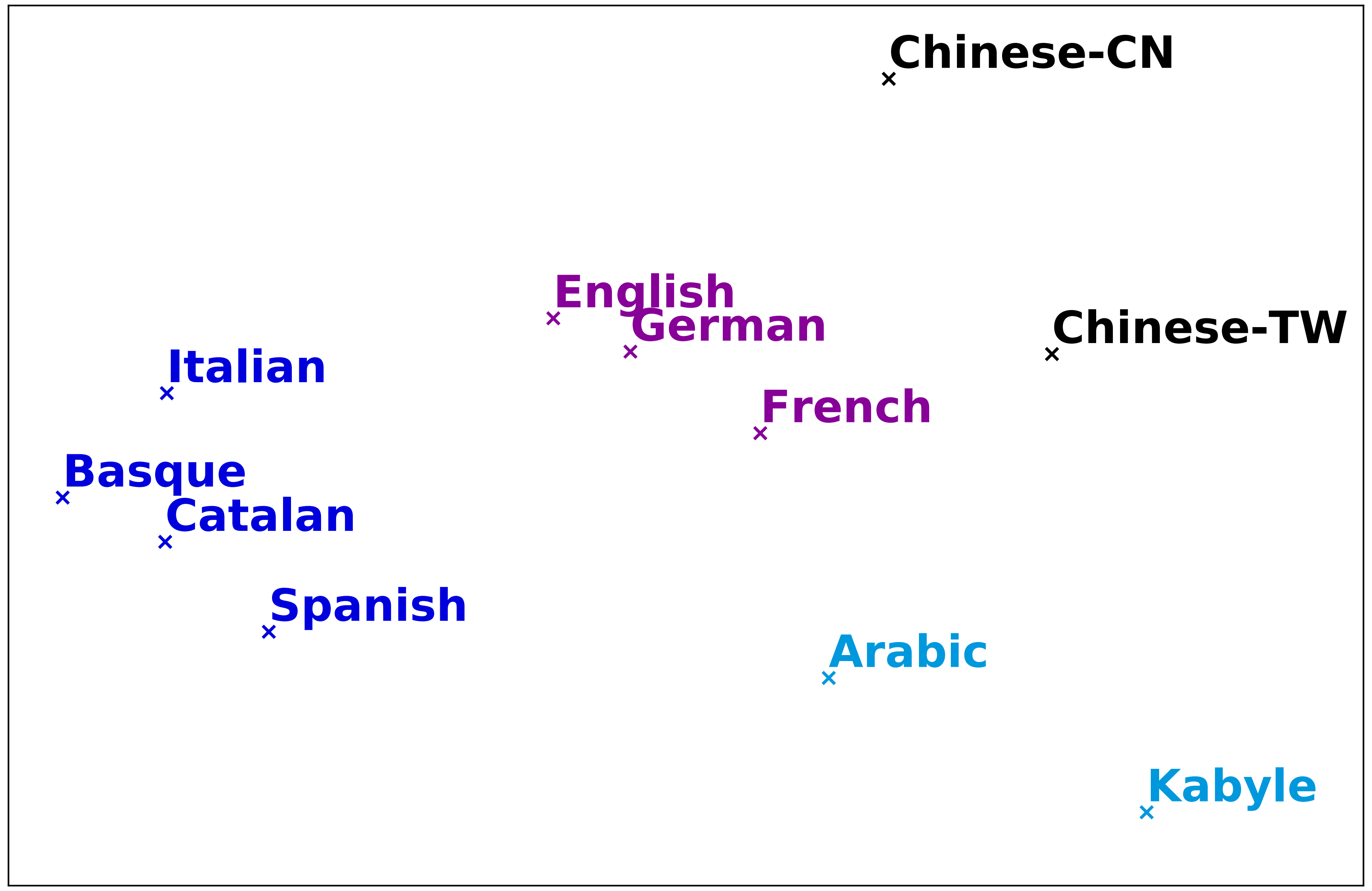
Similar higher-resource language data helps the most for low-resource language

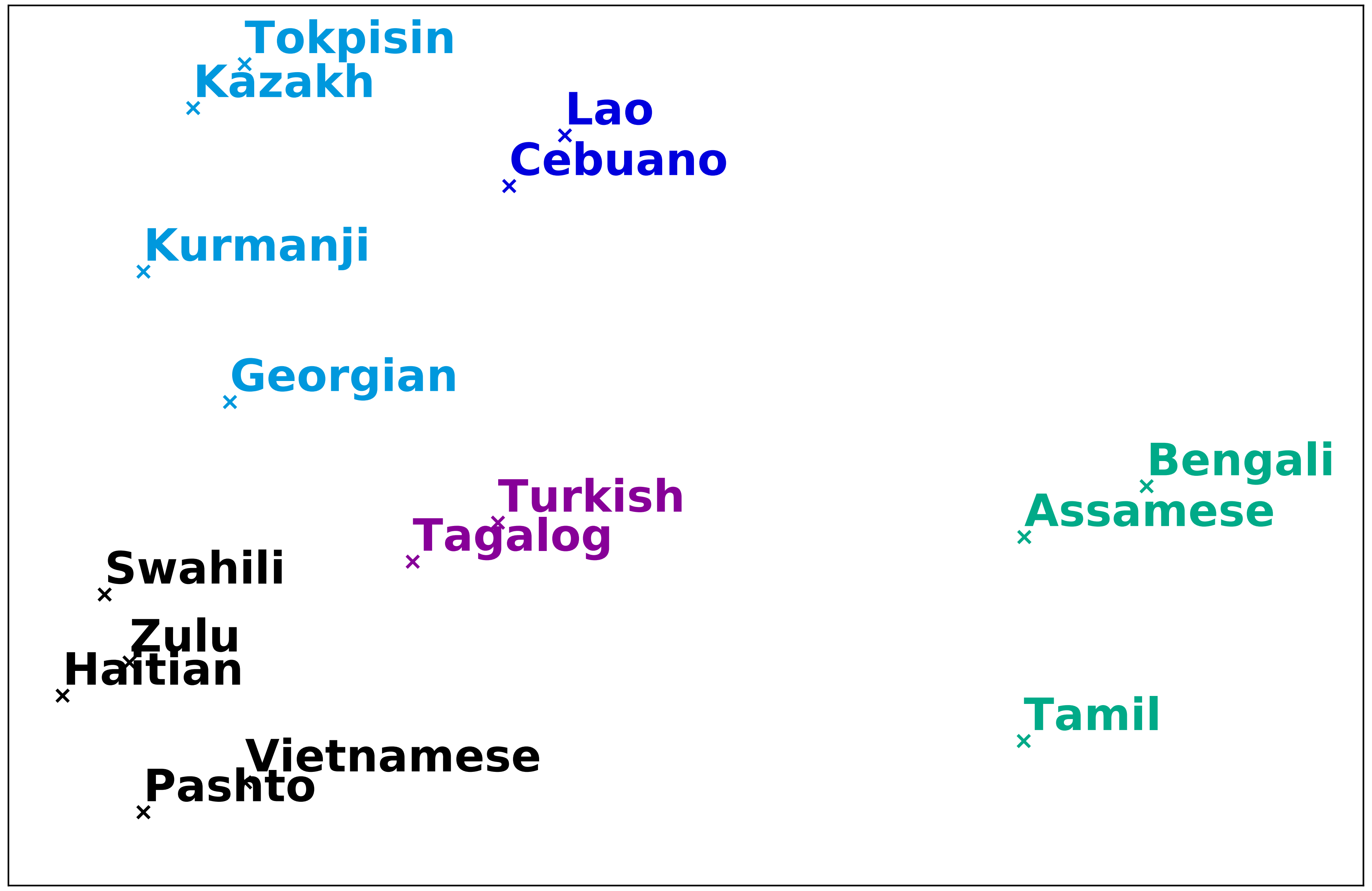


XLSR: Analysis of discrete latent speech representations

PCA visualization of latent discrete representations from the multilingual codebook

Similar languages tend to share discrete tokens and thus cluster together





Conclusion

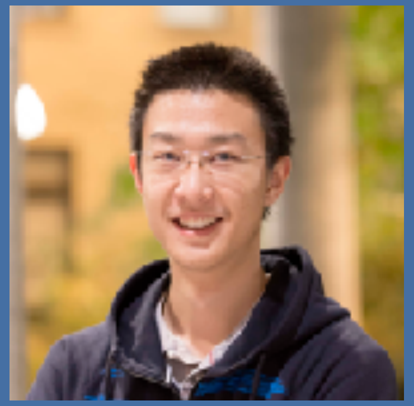
- For the first time, pre-training for speech works very well in both low-resource and high-resource setup.
- Cross-lingual training improves low-resource languages.
- Pre-training and self-training are complementary.
- Using only 10 minutes (48 utterances) of transcribed data rivals best system trained on 960h from 1 year ago.
- Code and models are available in the fairseq GitHub repo + Hugging Face.

Thank you


Alexei Baevski


Alexis Conneau


Steffen Schneider


Henry Zhou


Abdelrahman Mohamed


Anuroop Sriram


Naman Goyal


Wei-Ning Hsu


Michael Auli


Kritika Singh


Yatharth Saraf


Geoffrey Zweig


Qiantong Xu


Tatiana Likhomanenko


Paden Tomasello


Ronan Collobert


Gabriel Synnaeve