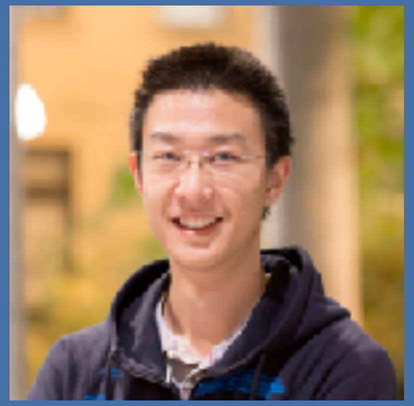
Self-supervised learning for speech processing

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

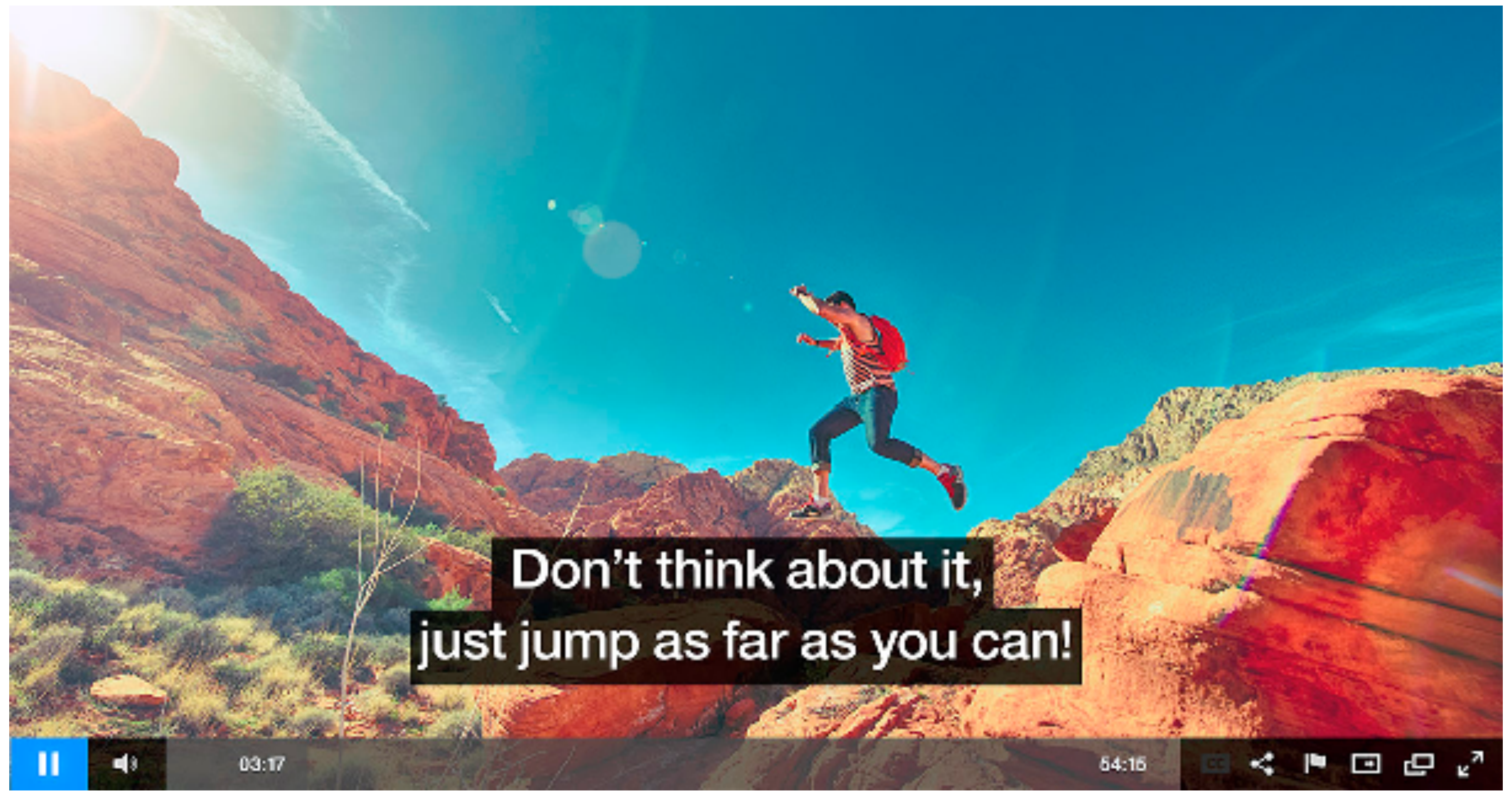

Tatiana Likhomanenko


Paden Tomasello


Ronan Collobert


Gabriel Synnaeve

Speech technology

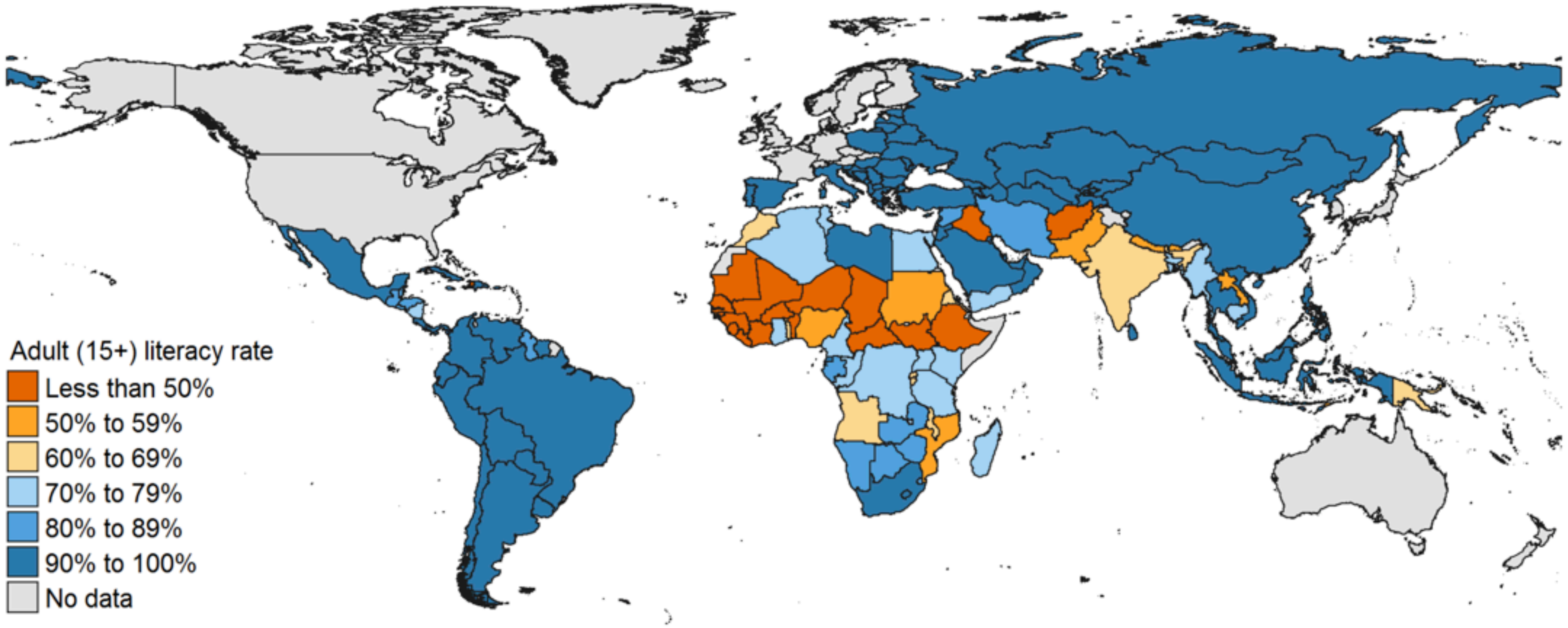


Video captioning

Mobile devices

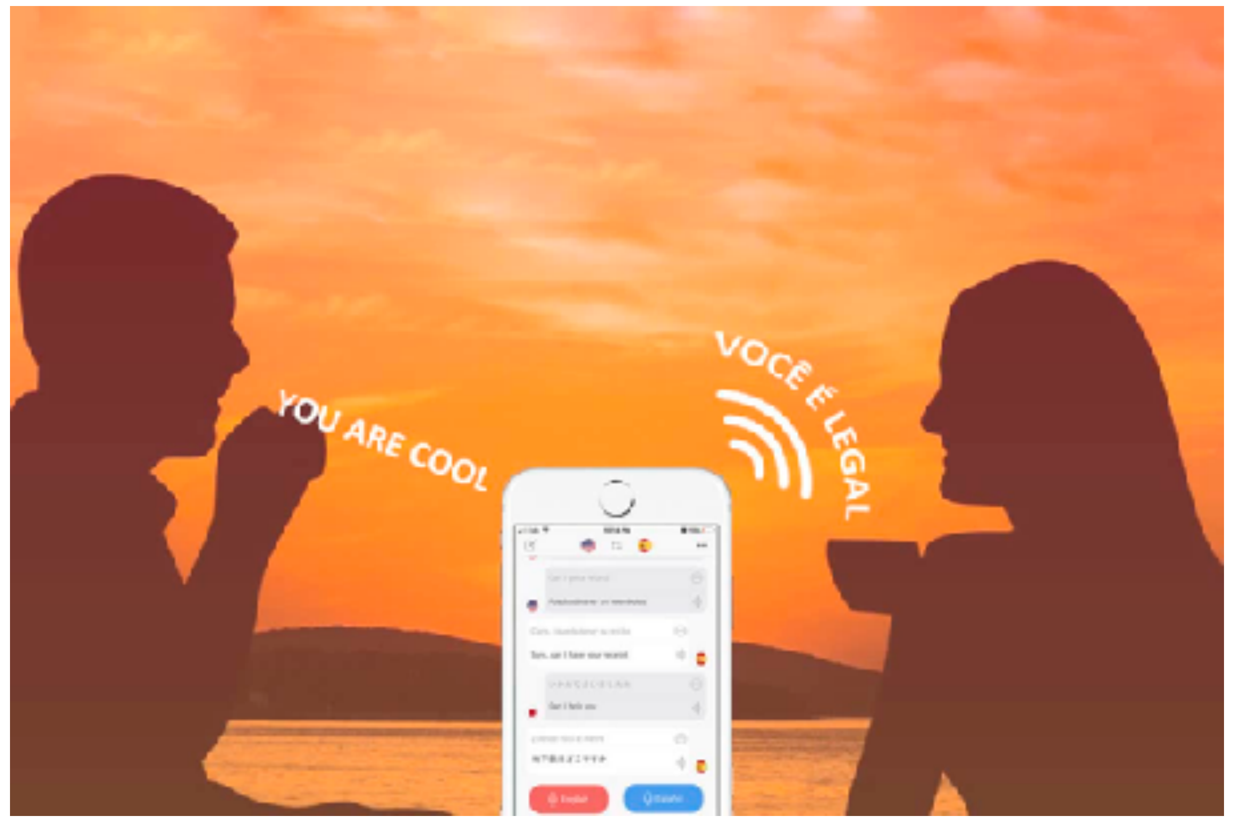
Home devices

Adult literacy rate by country, 2016



Speech applications

- Speech to text/speech recognition - dictation etc.
- Text to speech - reading out aloud
- Keyword spotting - "Hey Alexa/Portal"
- Speaker identification - is it your voice?
- Language identification
- Speech translation



Overview

- Traditional speech recognition

- Self-supervised learning for speech processing
  - wav2vec 2.0
  - Cross-lingual training
  - Completely unsupervised speech recognition

Traditional speech recognition

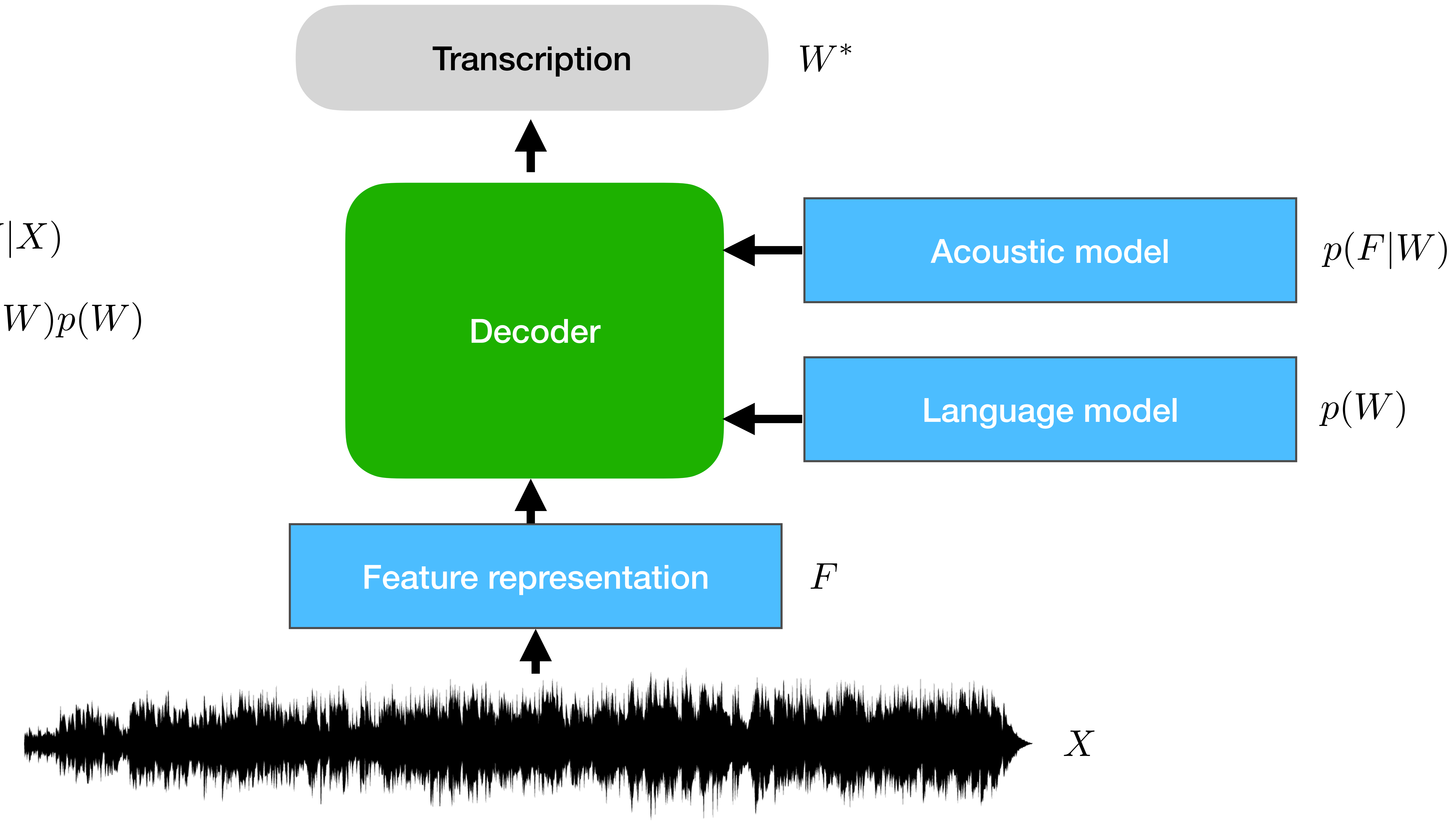
Speech recognition

I like black tea with milk

Traditional automatic speech recognition (ASR)

$$W^* = \arg\max_{W} p(W|X)$$

$$W^* = \arg\max_{W} p(F|W)p(W)$$



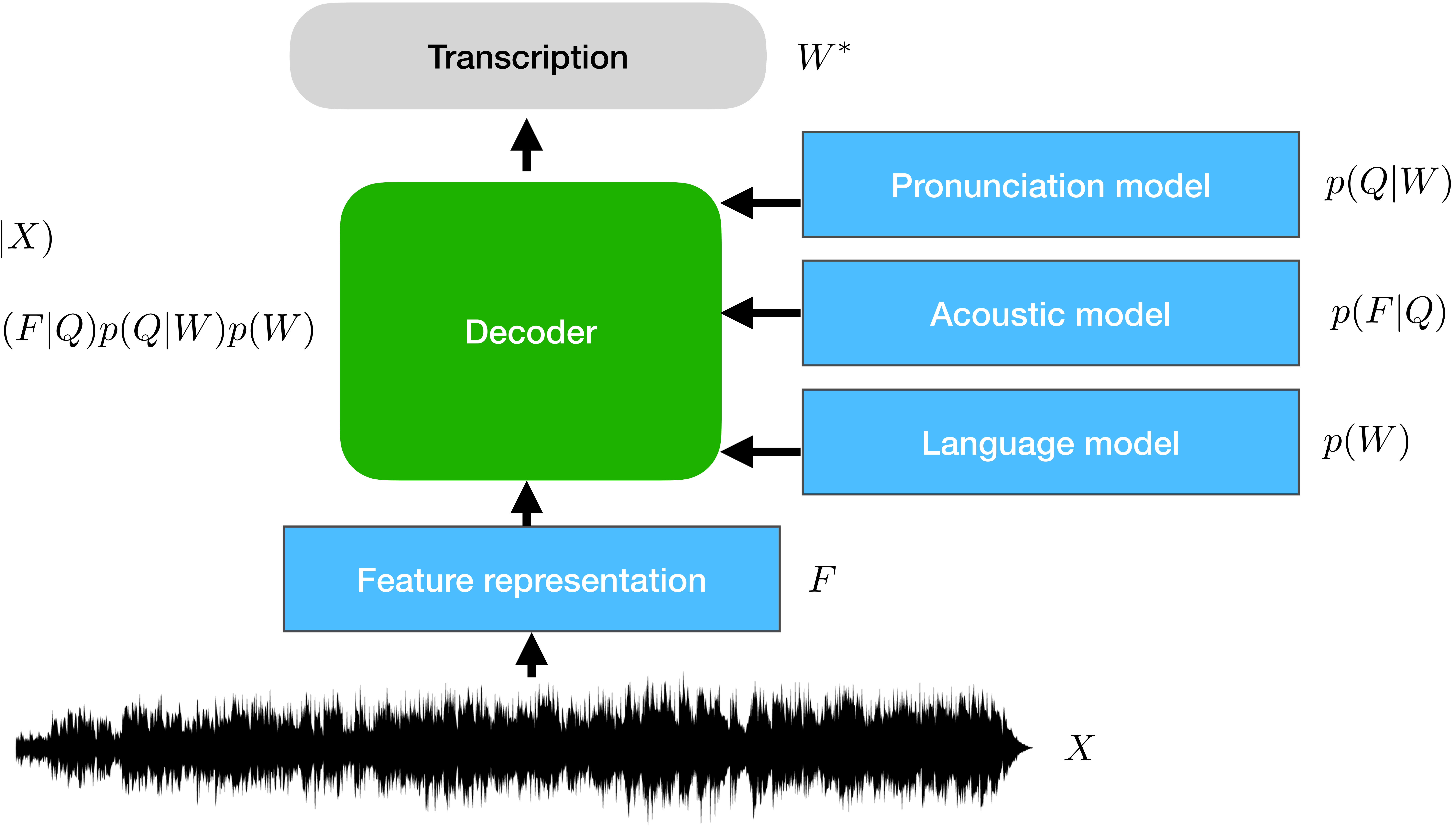
Traditional automatic speech recognition (ASR)

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

Traditional automatic speech recognition (ASR)

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

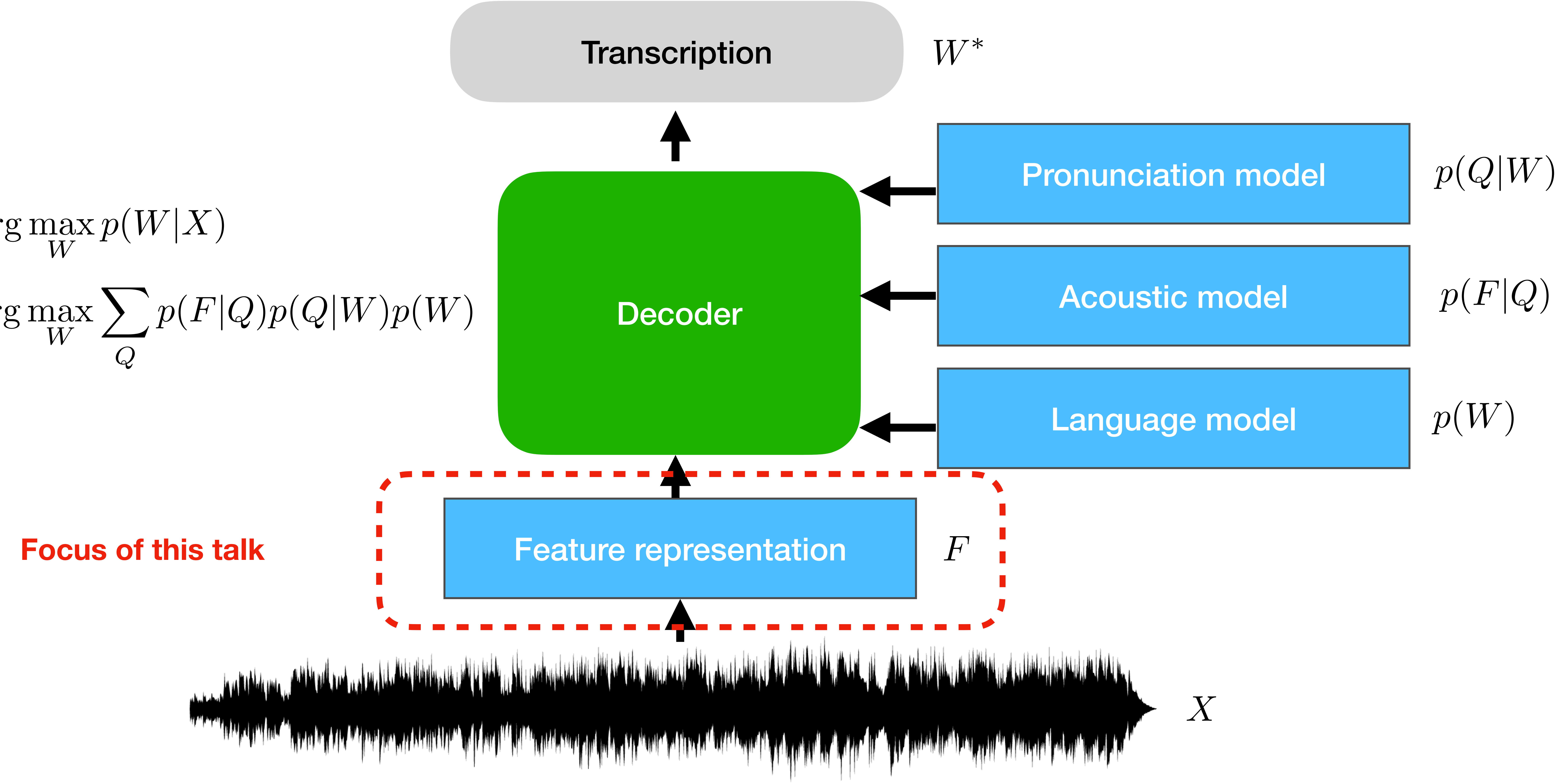
$$W^* = \arg\max_W \sum_Q p(F|Q)p(Q|W)p(W)$$



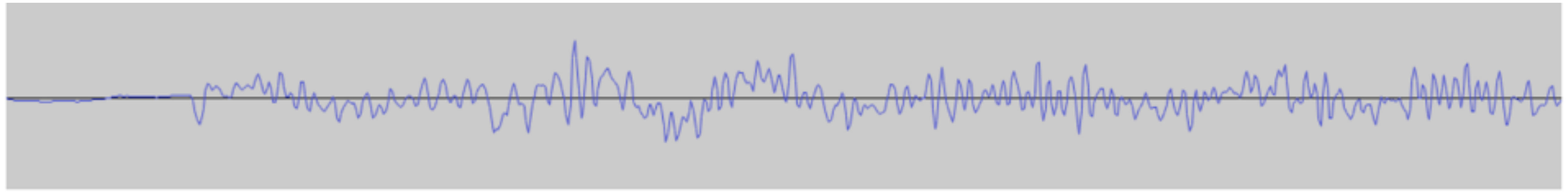
Traditional automatic speech recognition (ASR)

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

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



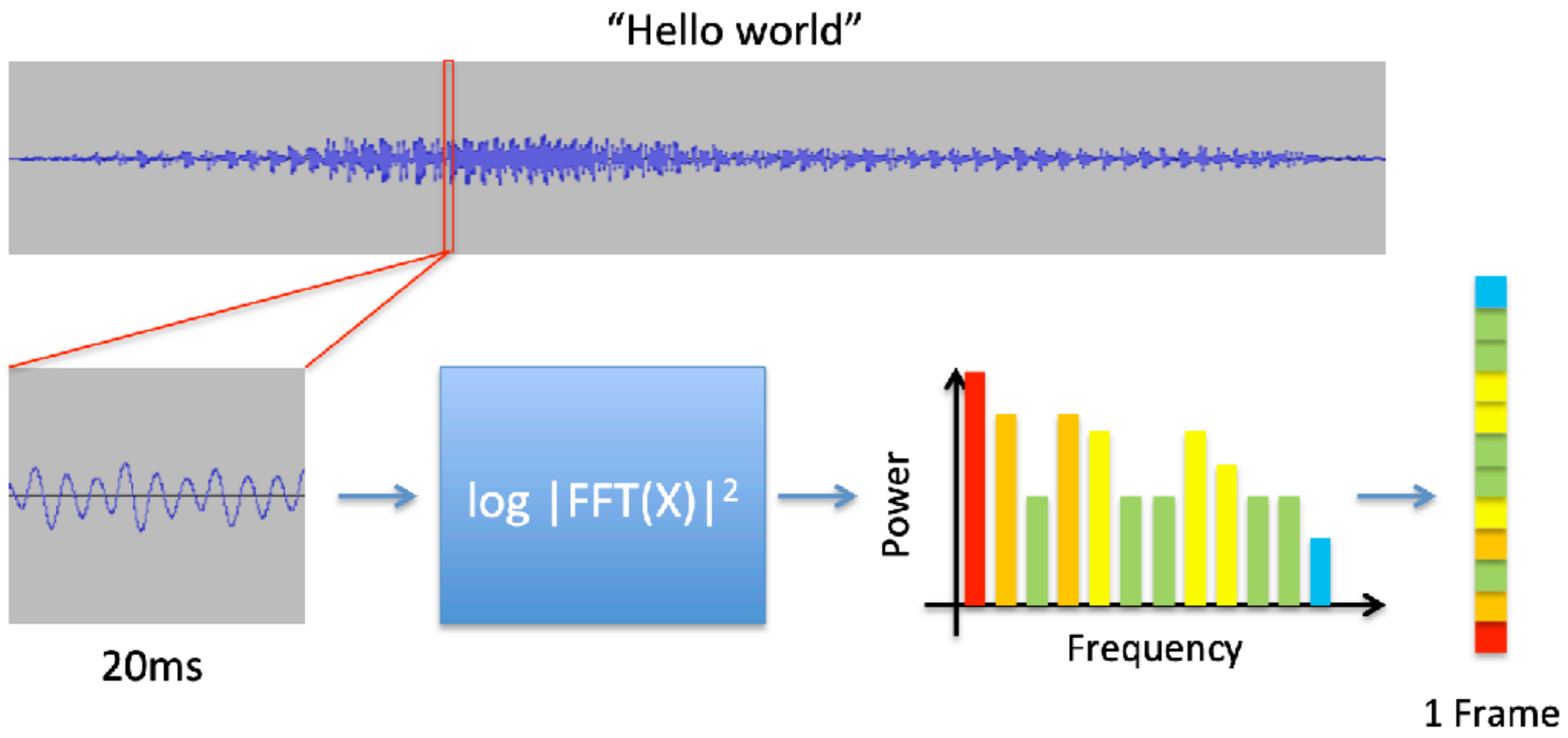
Feature representation



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

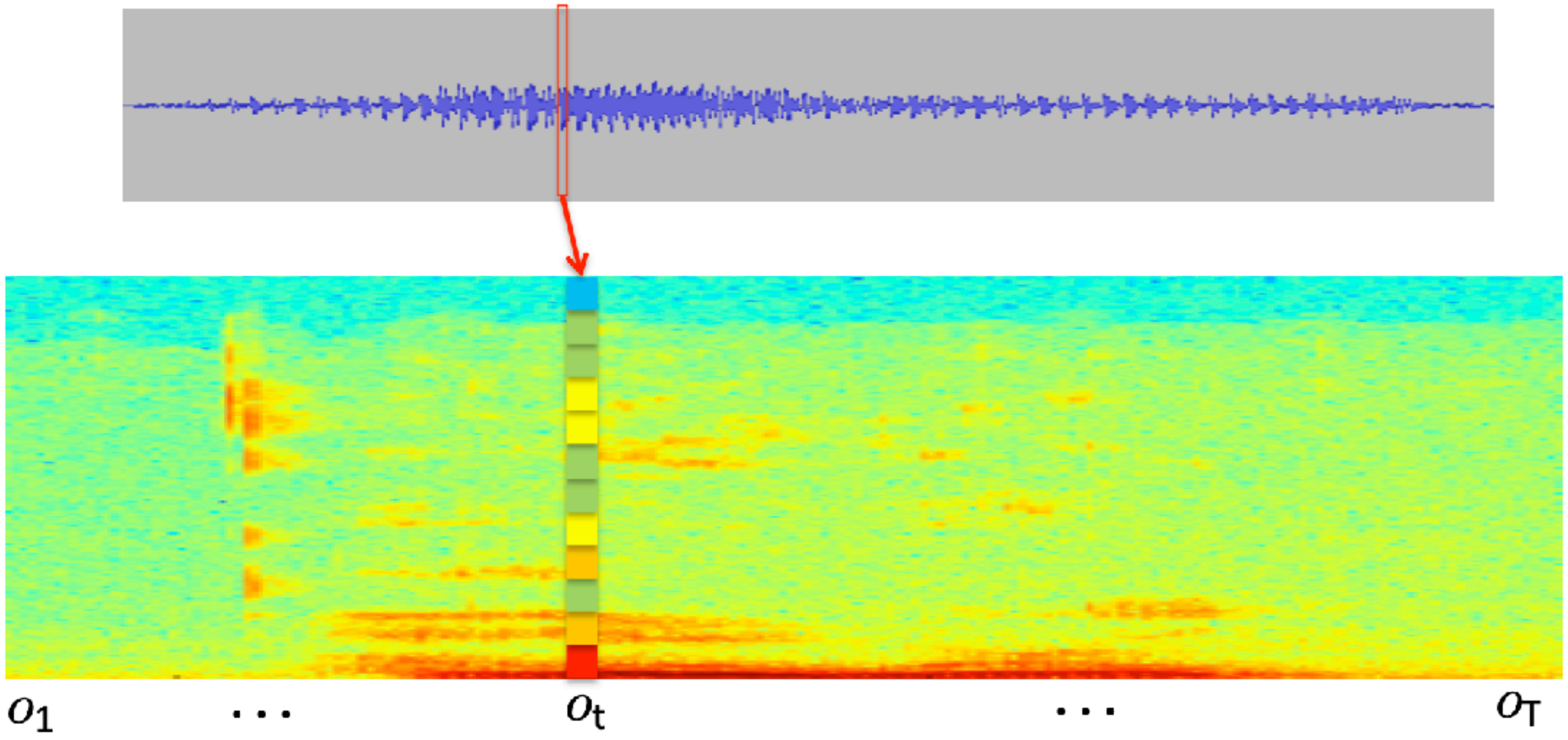
Spectrogram

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



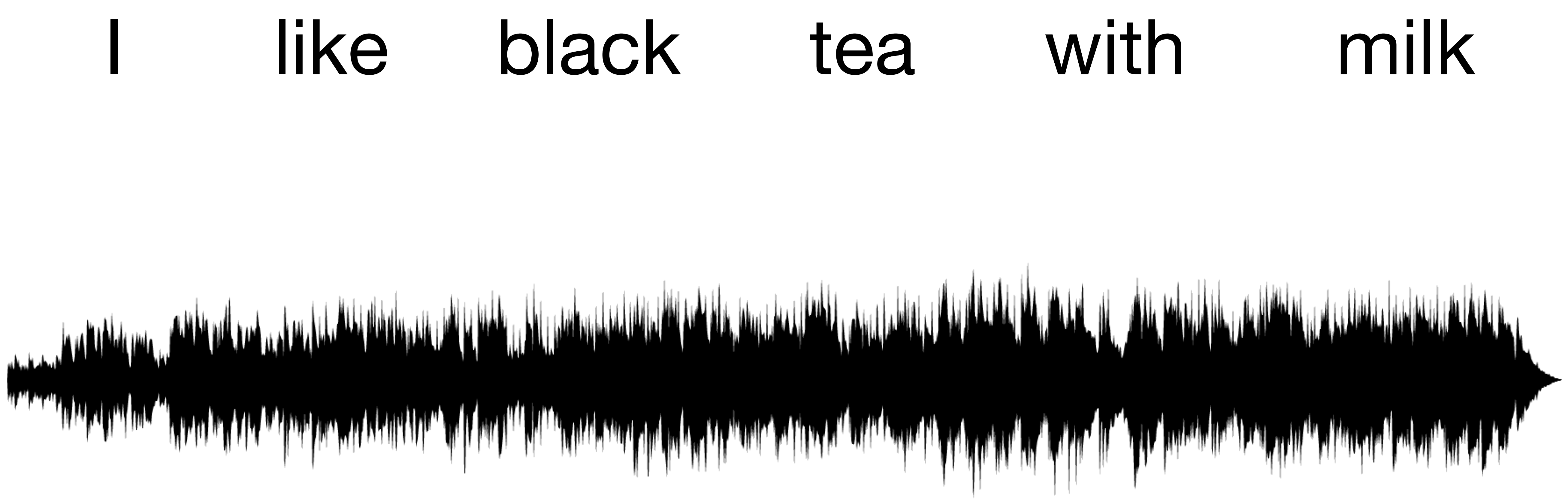
Spectrogram

- 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!

Supervised machine learning

( , cat)

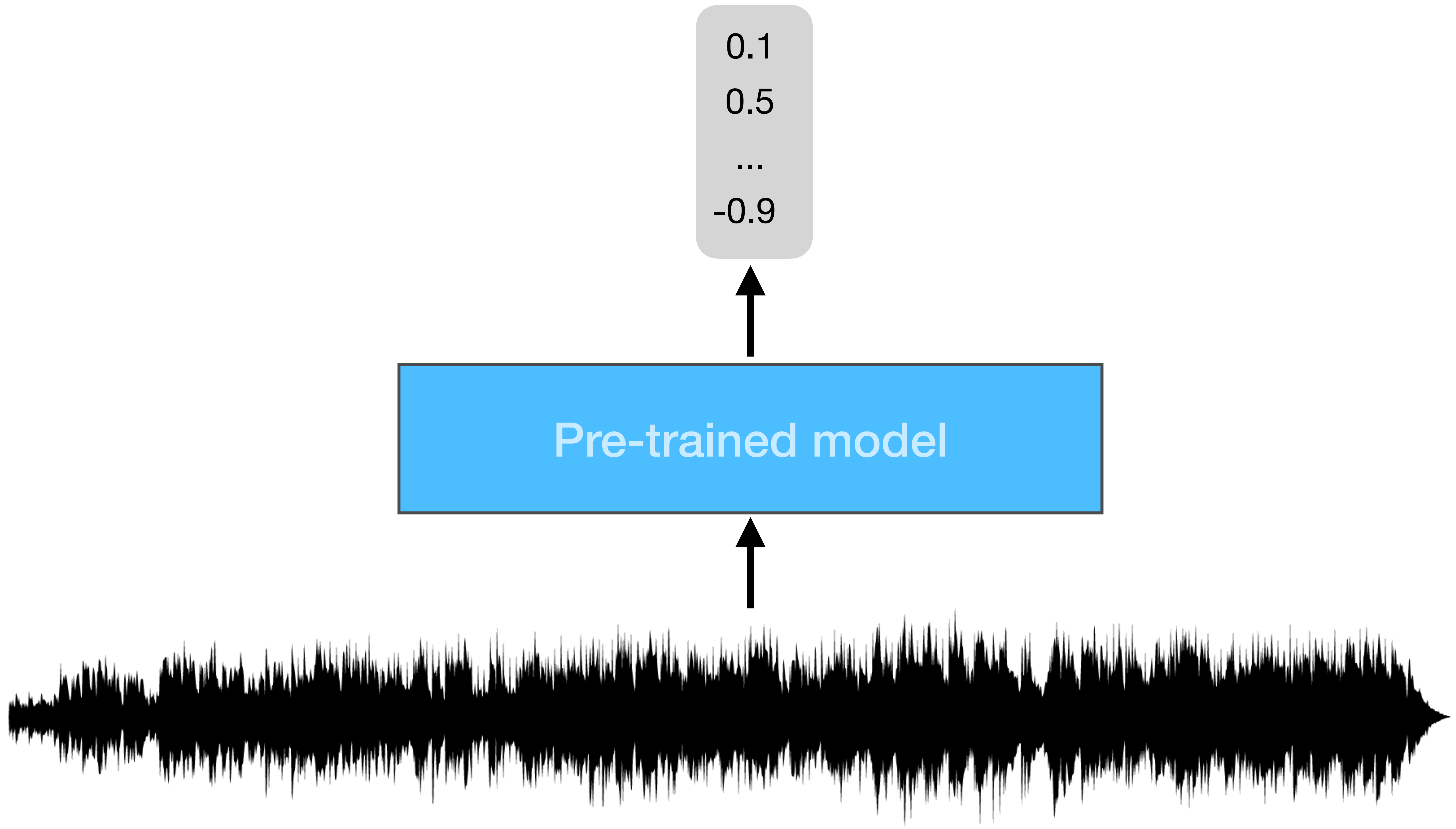
Not how humans learn!

potential train/test mismatch 

Need to annotate lots of data!

Supervised machine learning

Learning good representations of audio data
from unlabeled audio


0.1
0.5
...
-0.9
Pre-trained model

Speech recognition

I like tea

0.1
0.5
...
-0.9

Pre-trained model

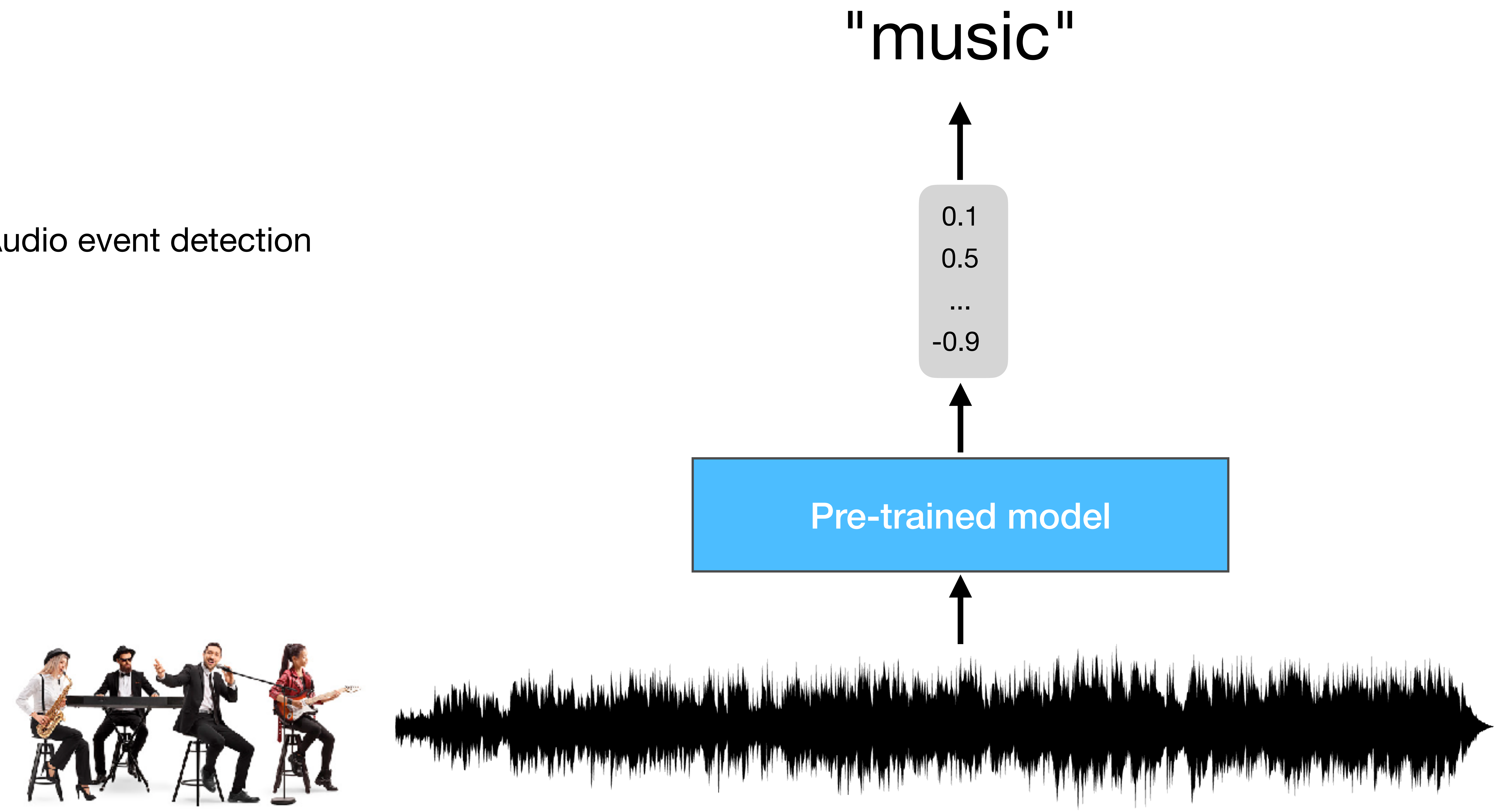
Speech translation

Ich mag Tee

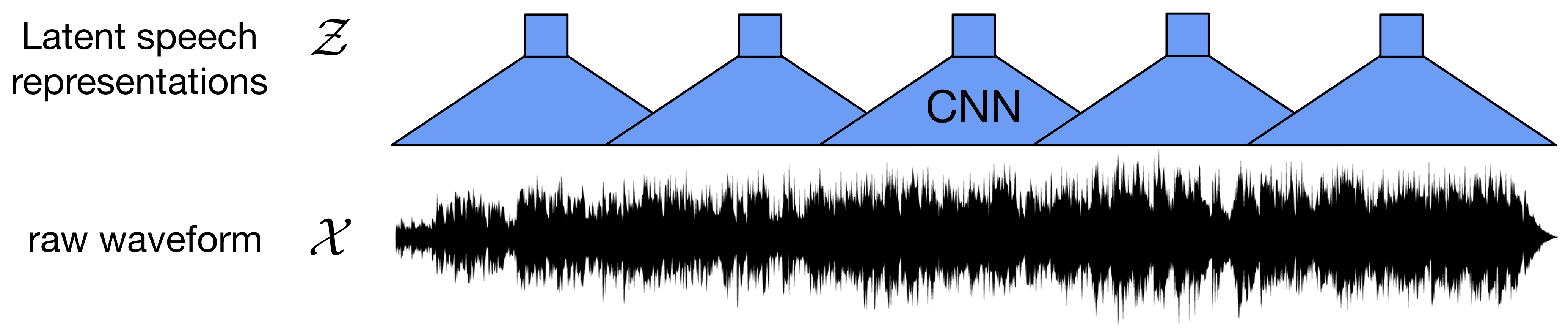
0.1
0.5
...
-0.9

Pre-trained model


Audio event detection

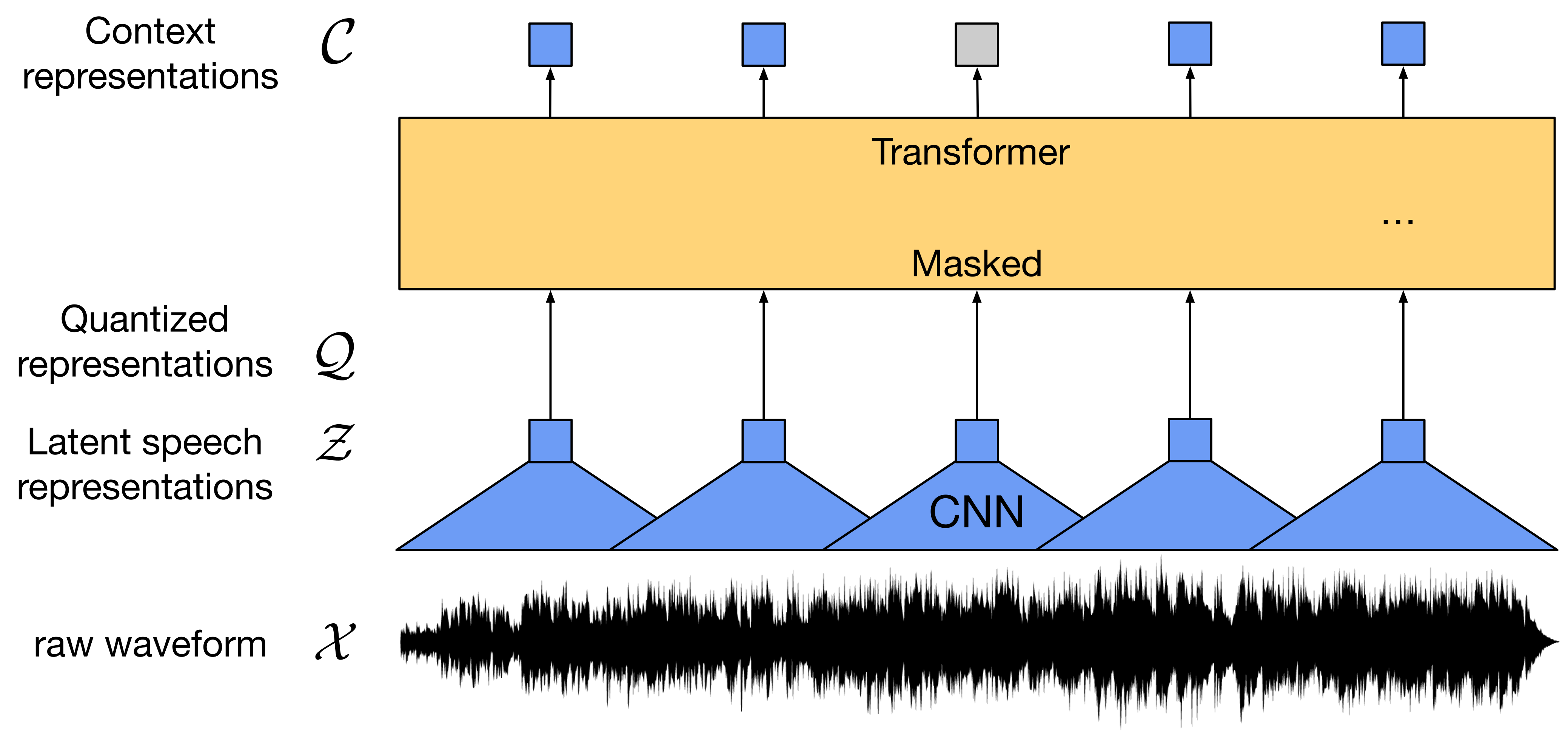


wav2vec 2.0



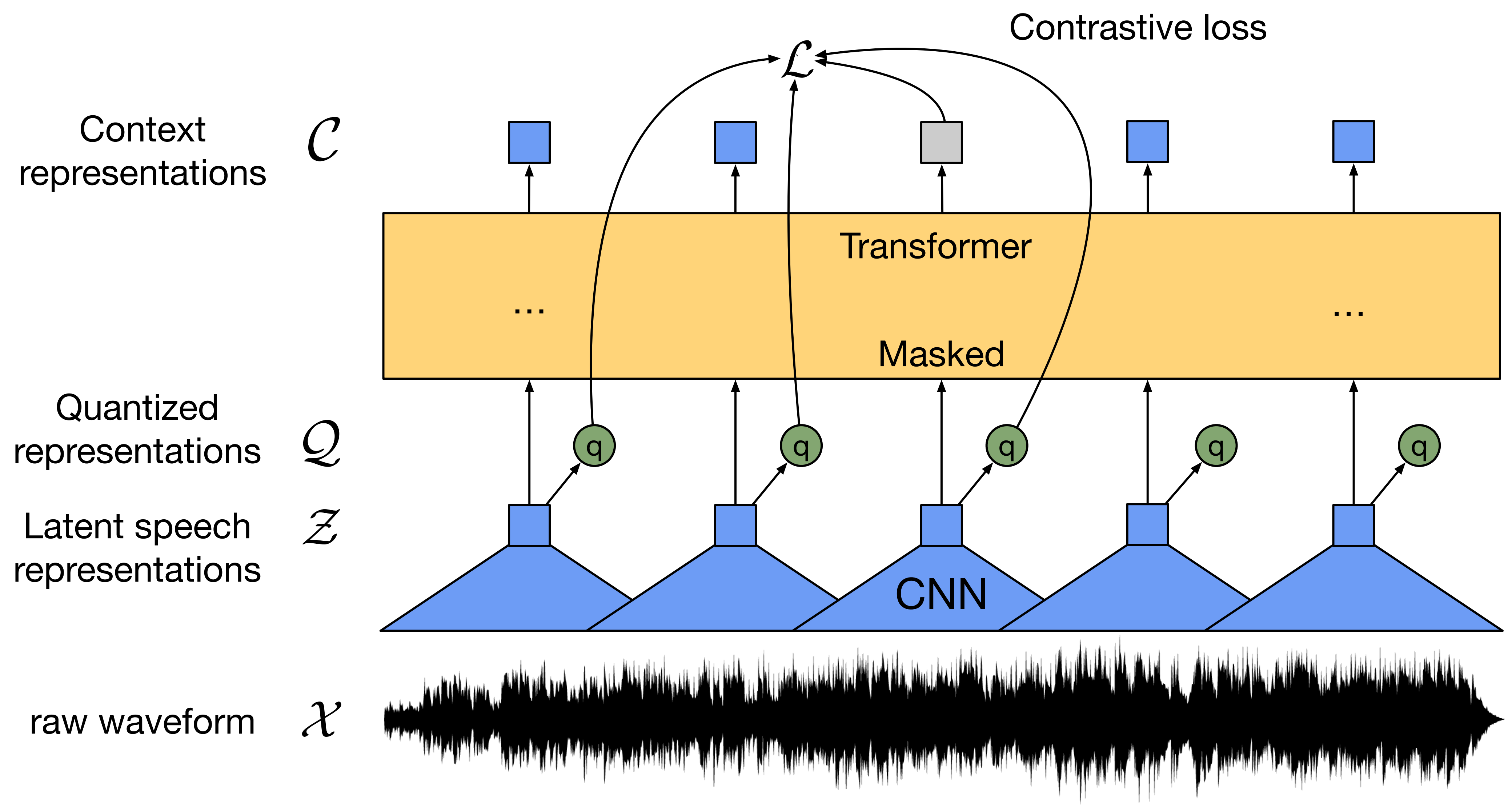
- Masked prediction with transformer, bi-directional contextualized representations (similar to BERT).
- But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.
- Learning task: Joint VQ & context representation learning.
- Contrast true quantized latent with distractor latents.

wav2vec 2.0



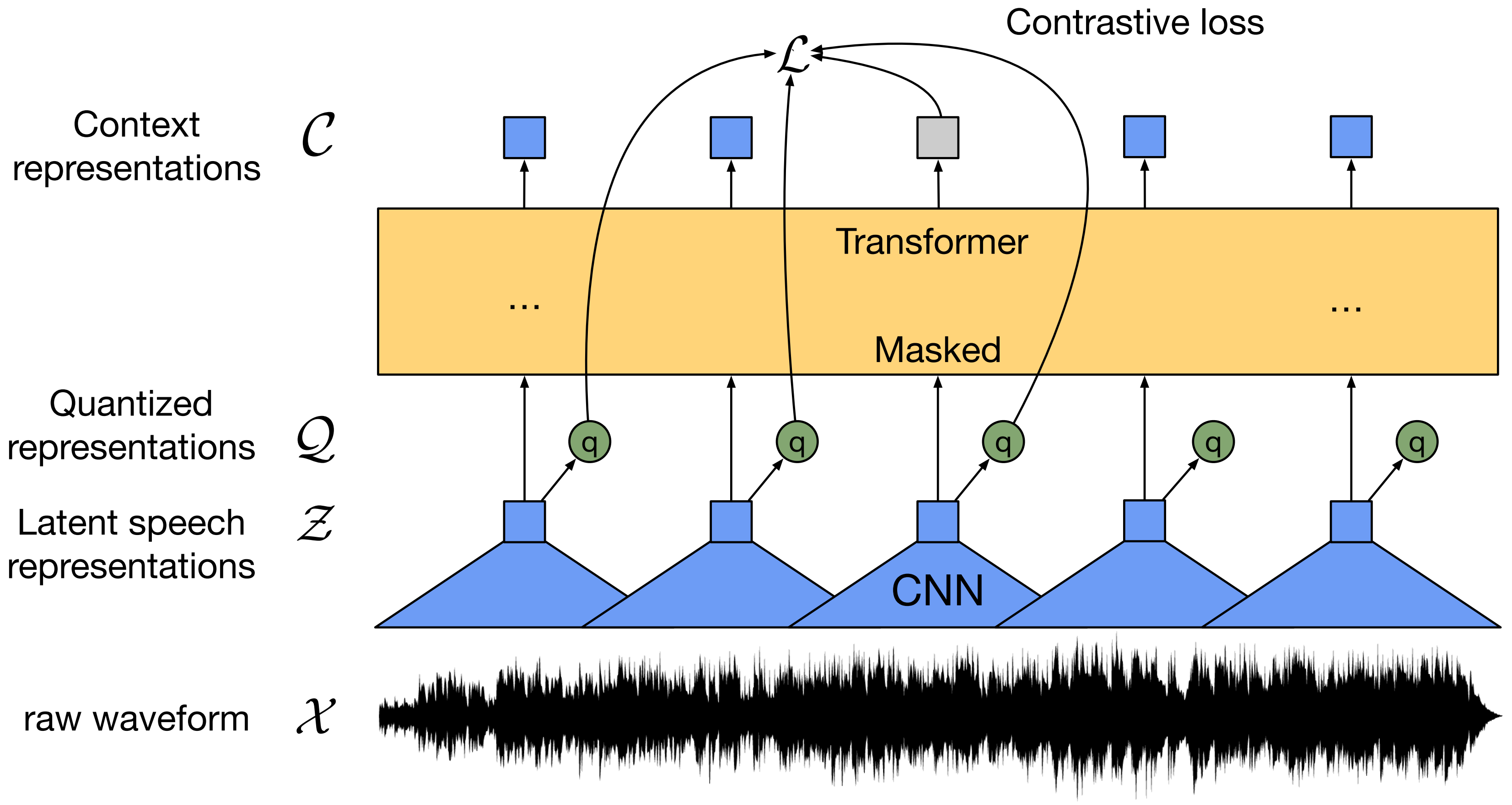
- Masked prediction with transformer, bi-directional contextualized representations (similar to BERT).
- But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.
- Learning task: Joint VQ & context representation learning.
- Contrast true quantized latent with distractor latents.

wav2vec 2.0



- Masked prediction with transformer, bi-directional contextualized representations (similar to BERT).
- But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.
- Learning task: Joint VQ & context representation learning.
- Contrast true quantized latent with distractor latents.

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

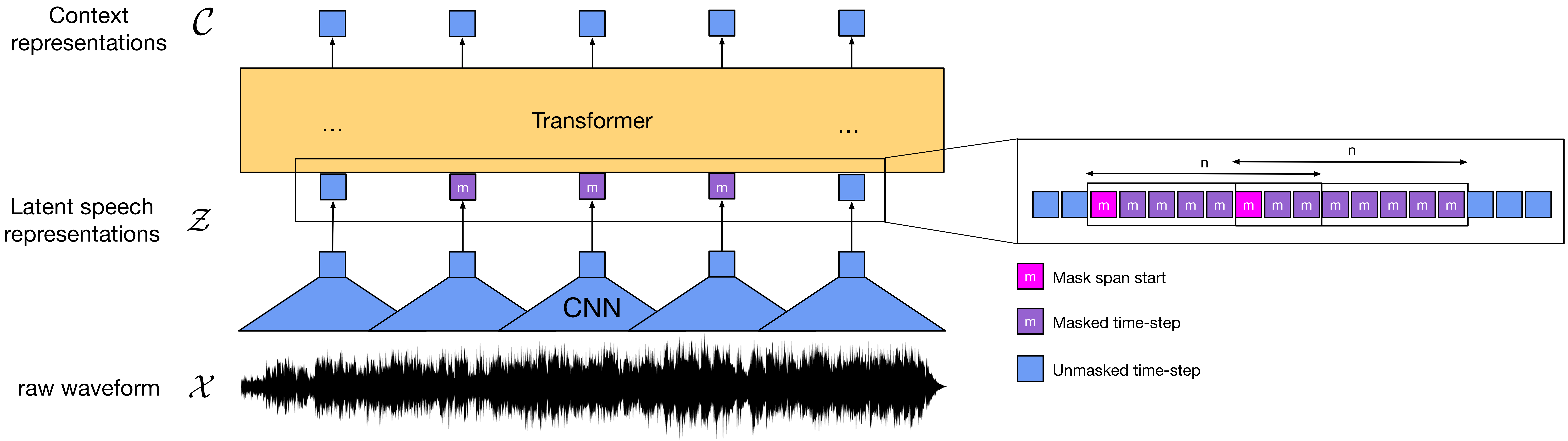
Codebook diversity penalty to encourage more codes to be used

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

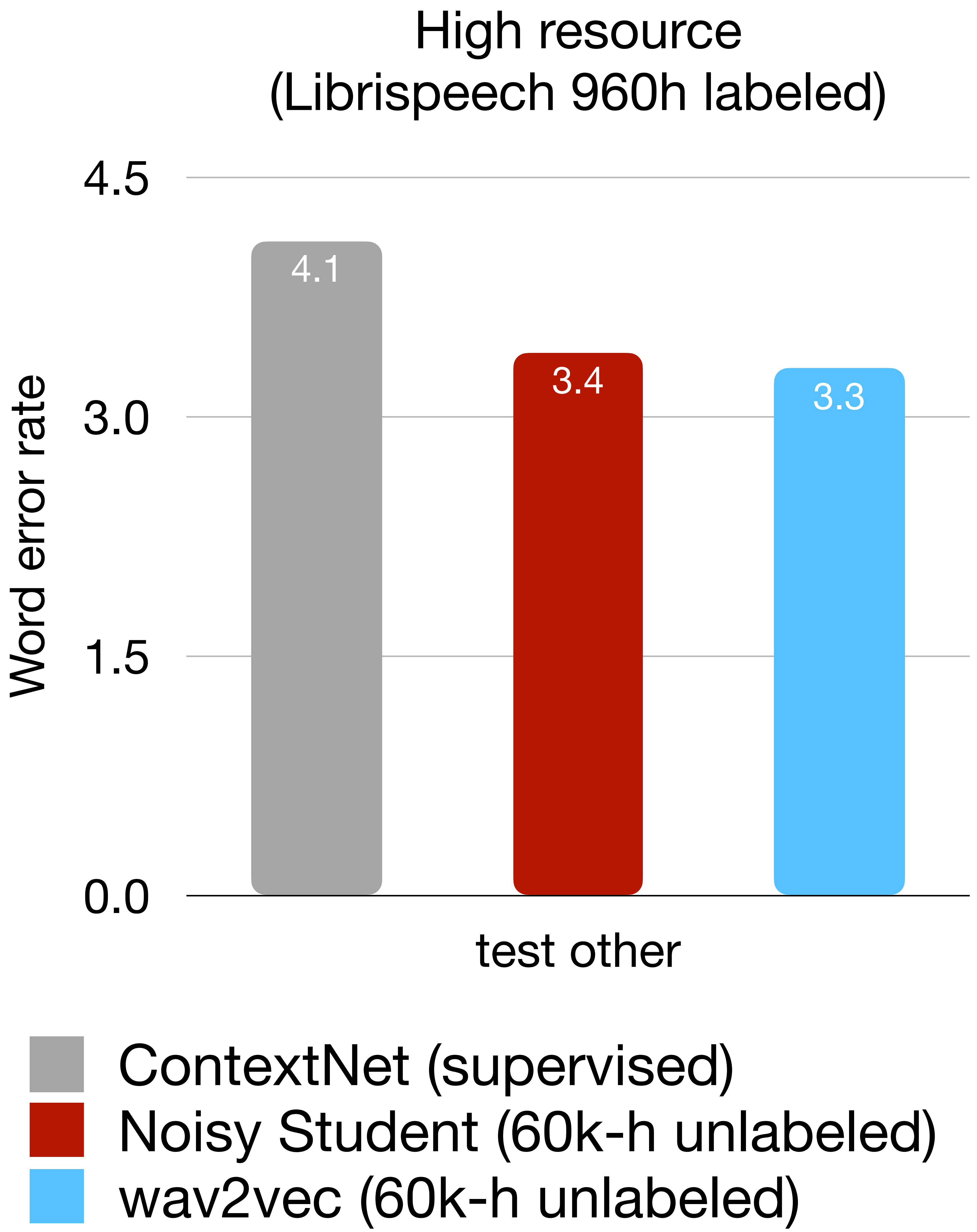
- For a 15s sample, ~49% of the time-steps masked with an average span length of ~300ms



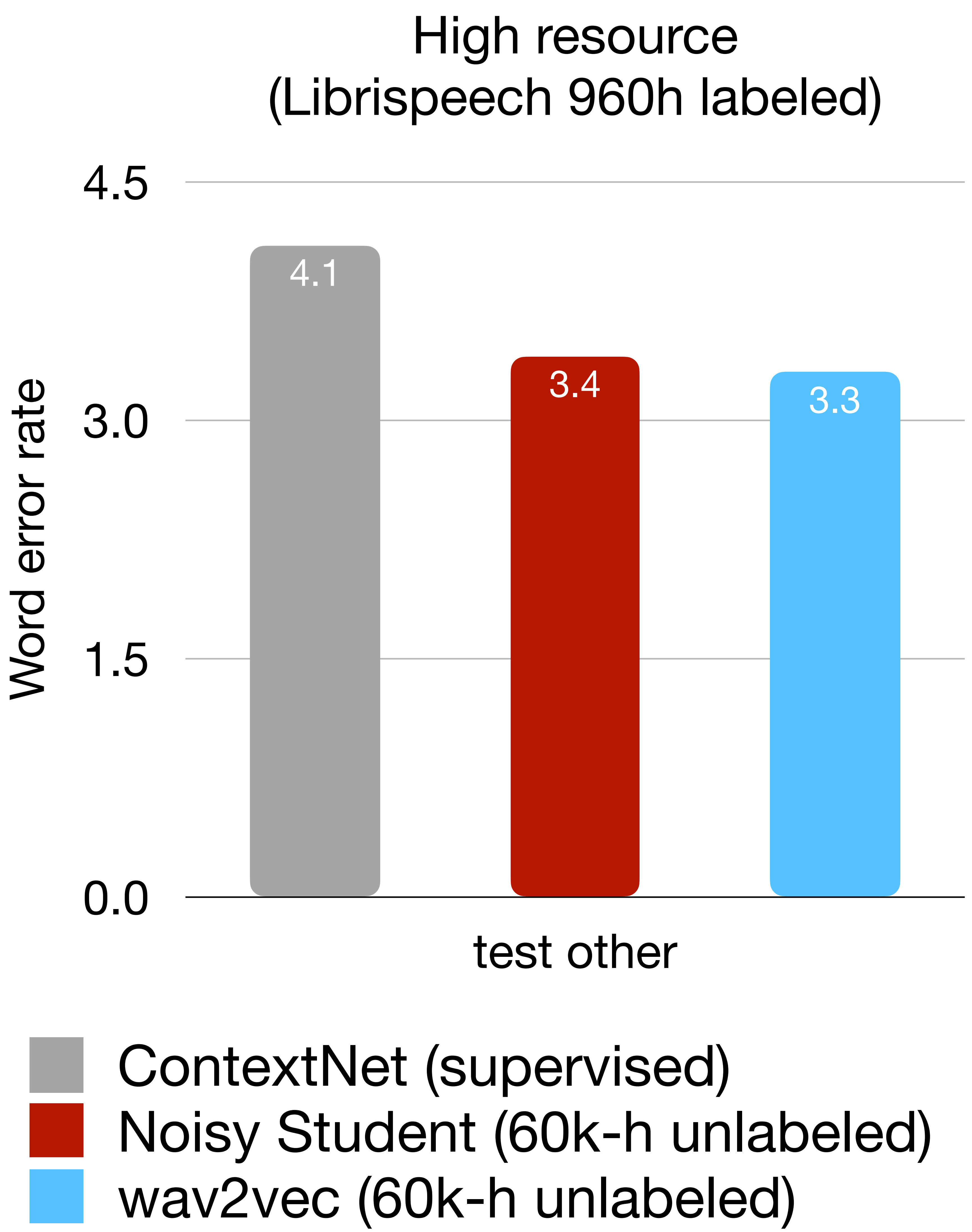
Fine-tuning

- 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

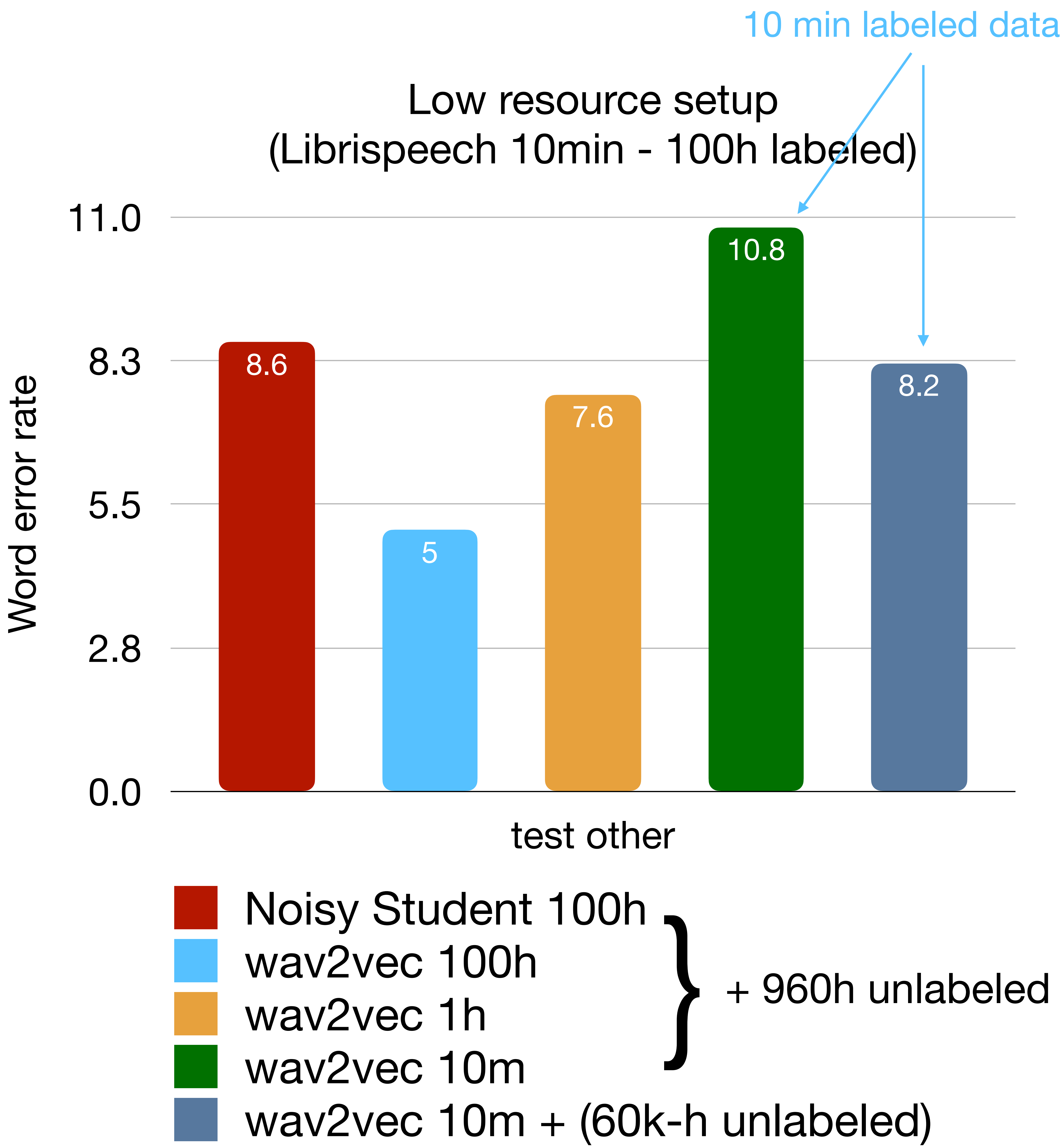
Results


High resource
(Librispeech 960h labeled)
4.5
3.0
1.5
0.0
Word error rate
4.1
3.4
3.3
test other
ContextNet (supervised)
Noisy Student (60k-h unlabeled)
wav2vec (60k-h unlabeled)

Results

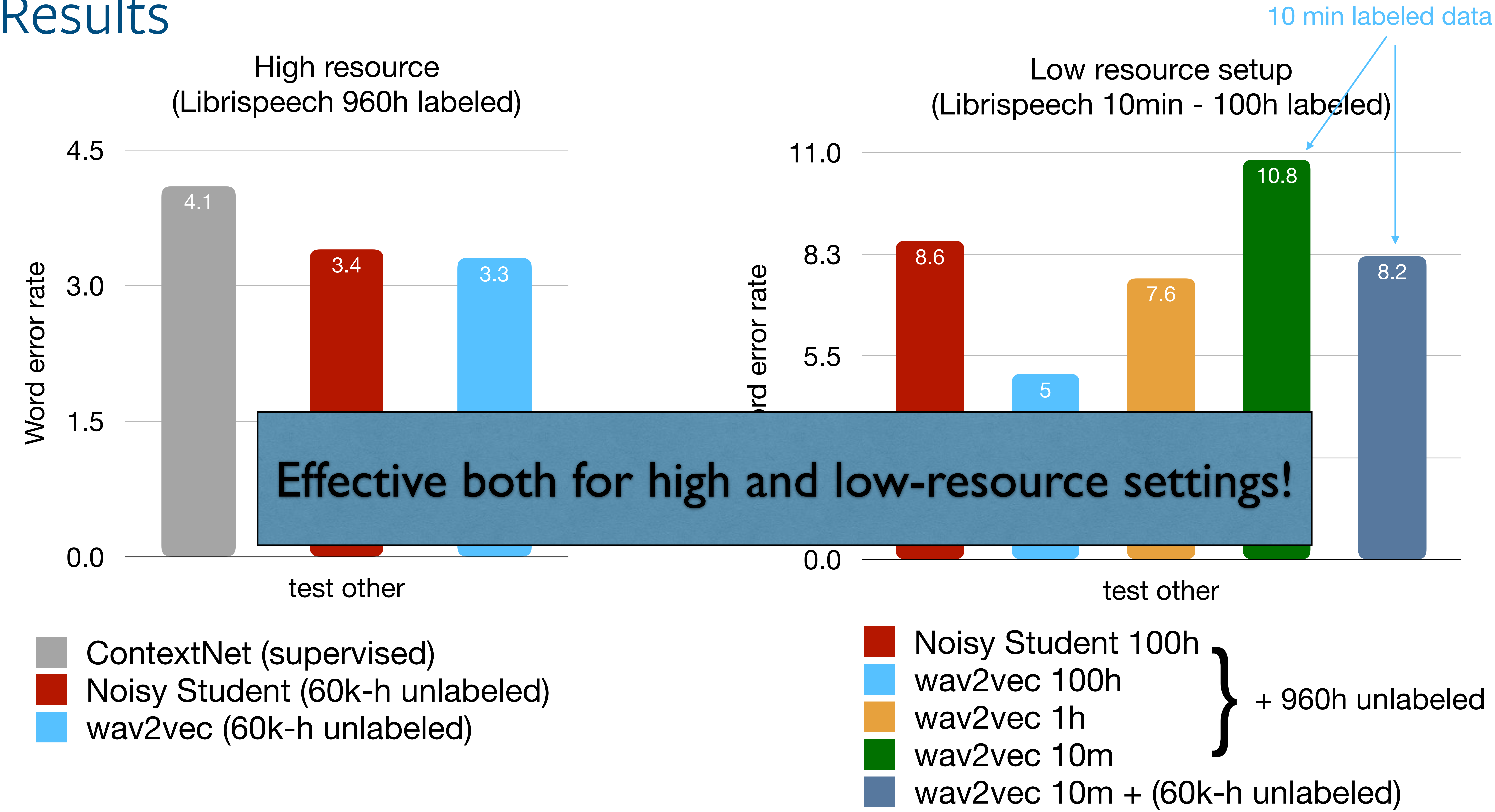


High resource
(Librispeech 960h labeled)
Word error rate
4.5
3.0
1.5
0.0
4.1
3.4
3.3
test other
ContextNet (supervised)
Noisy Student (60k-h unlabeled)
wav2vec (60k-h unlabeled)



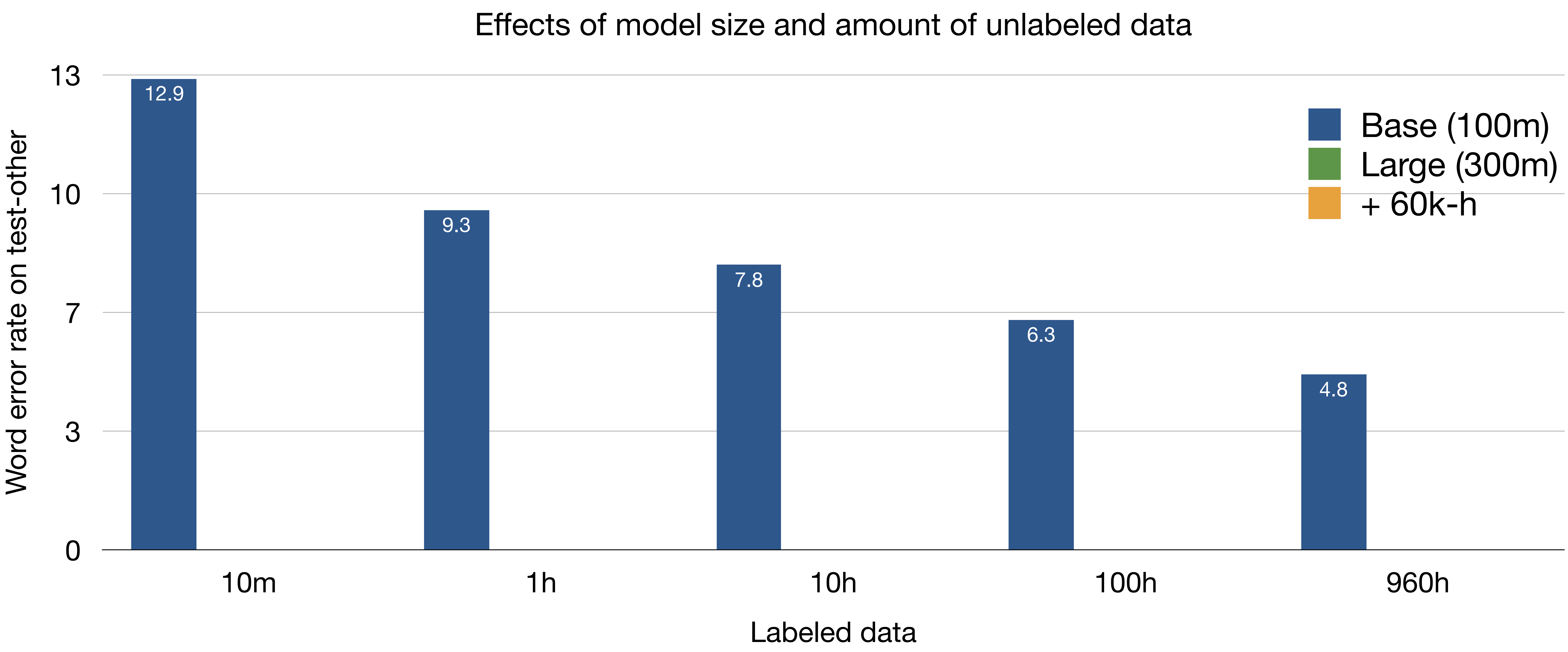
10 min labeled data
Low resource setup
(Librispeech 10min - 100h labeled)
Word error rate
11.0
8.3
5.5
2.8
0.0
8.6
5
7.6
10.8
8.2
test other
Noisy Student 100h
wav2vec 100h
wav2vec 1h
wav2vec 10m
+ 960h unlabeled
wav2vec 10m + (60k-h unlabeled)

Results



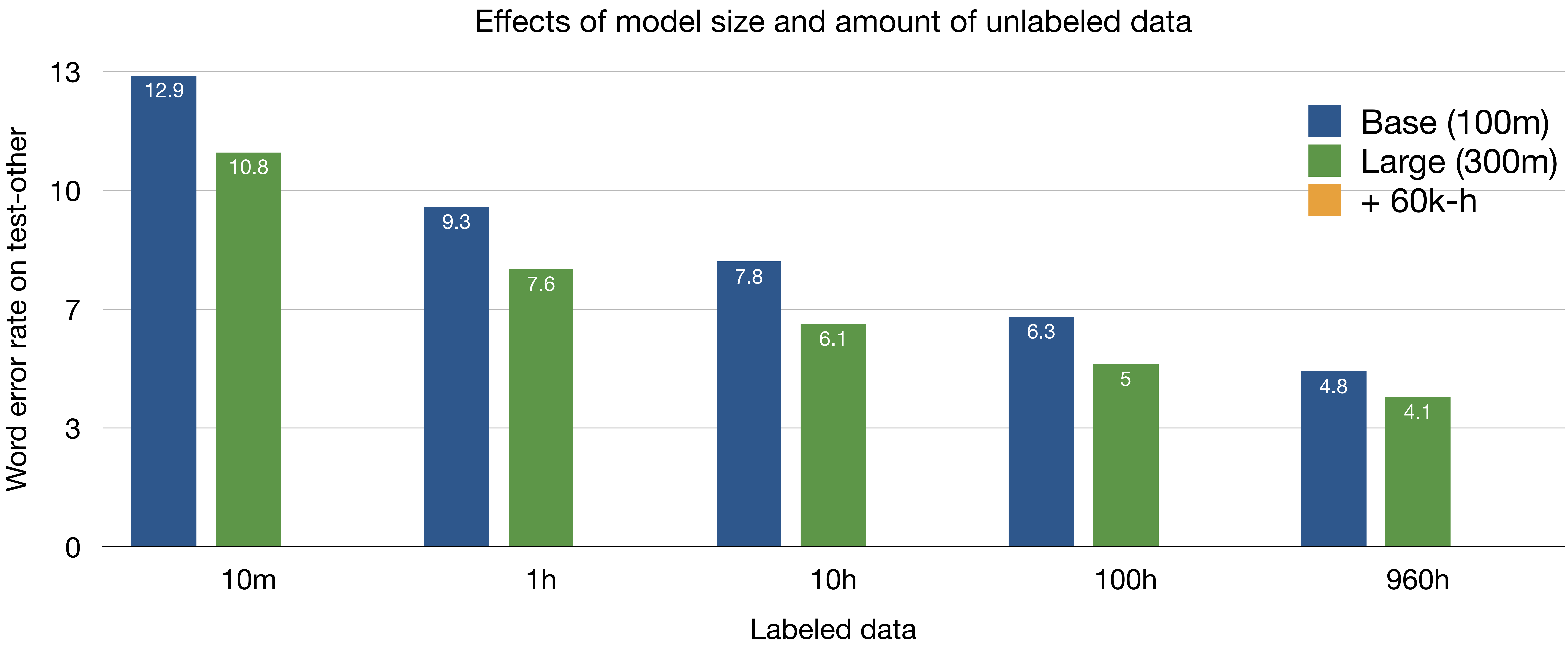
Effective both for high and low-resource settings!

Results

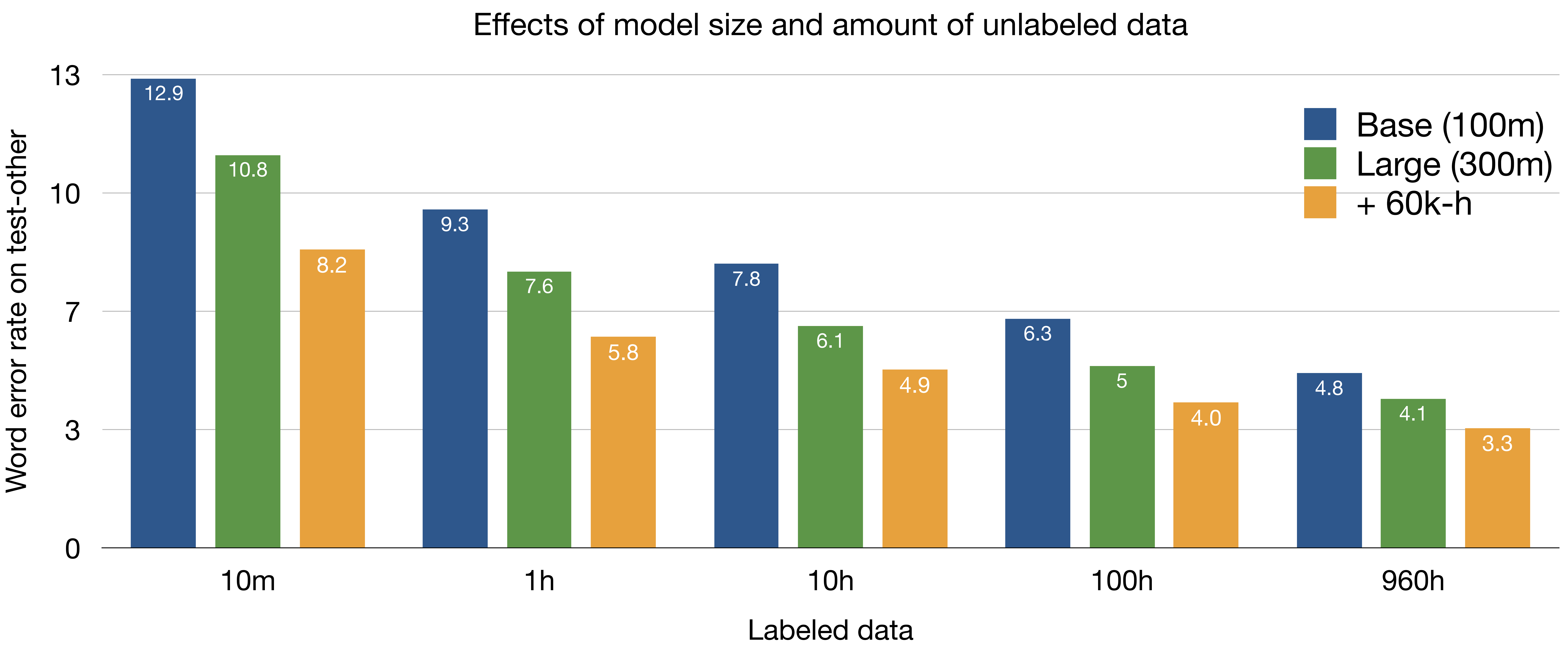


Effects of model size and amount of unlabeled data
Word error rate on test-other
13
10
7
3
0
12.9
9.3
7.8
6.3
4.8
10m
1h
10h
100h
960h
Labeled data
Base (100m)
Large (300m)
+ 60k-h

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

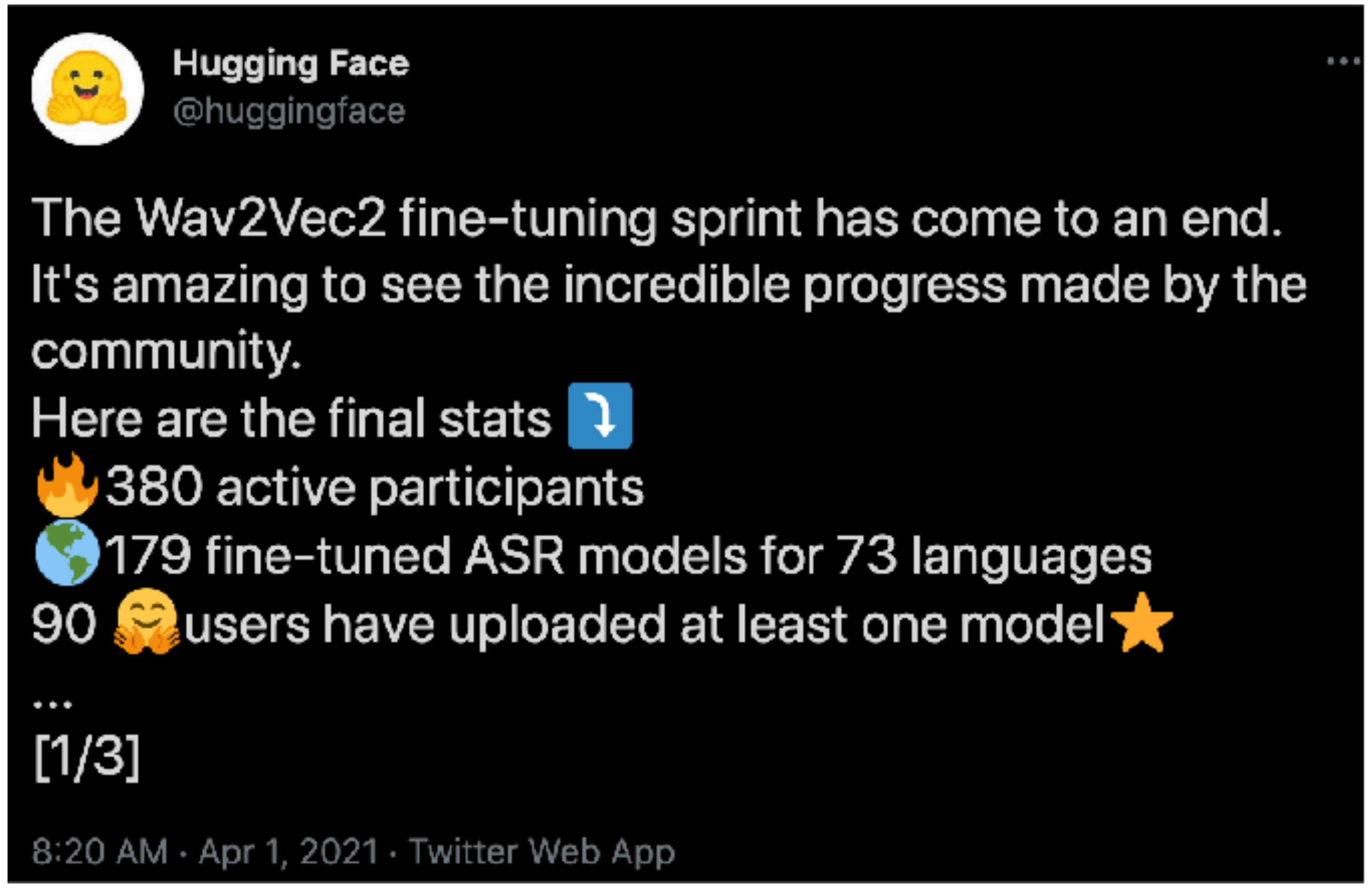
Examples (10 min labeled data)

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 Hugging Face

- Hugging Face is a popular NLP model zoo
- Hugging Face community fine-tuned our models to do speech recognition in 73 languages.



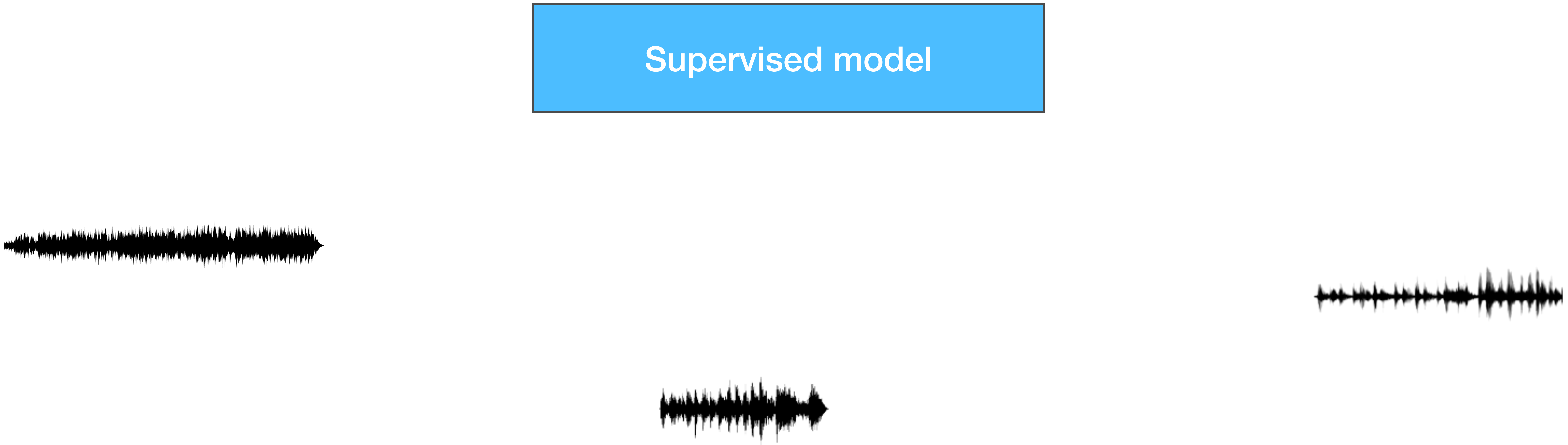
Pre-training and self-training

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

Supervised model

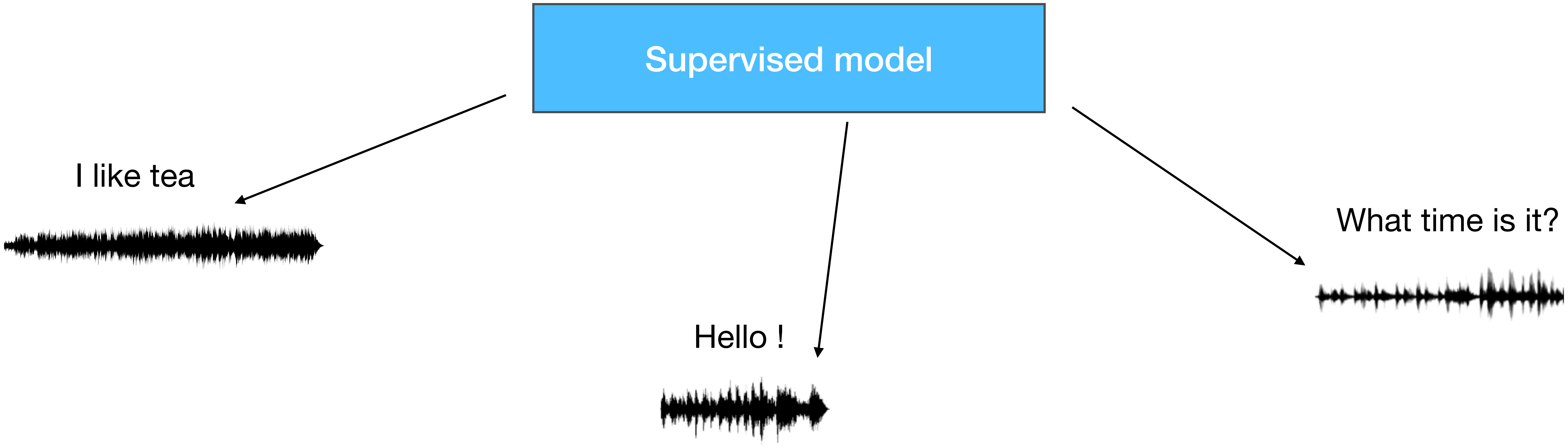
Pre-training and self-training

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



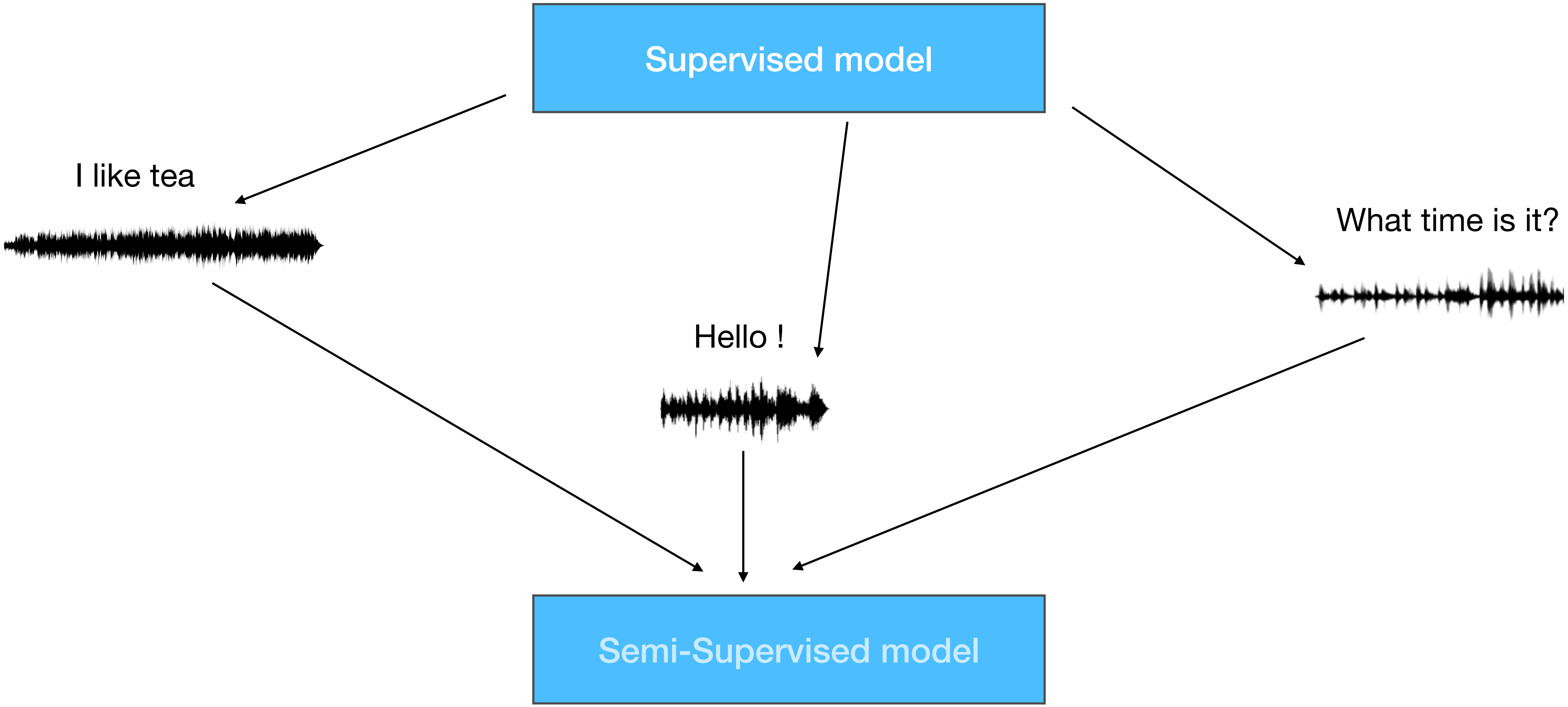
Pre-training and self-training

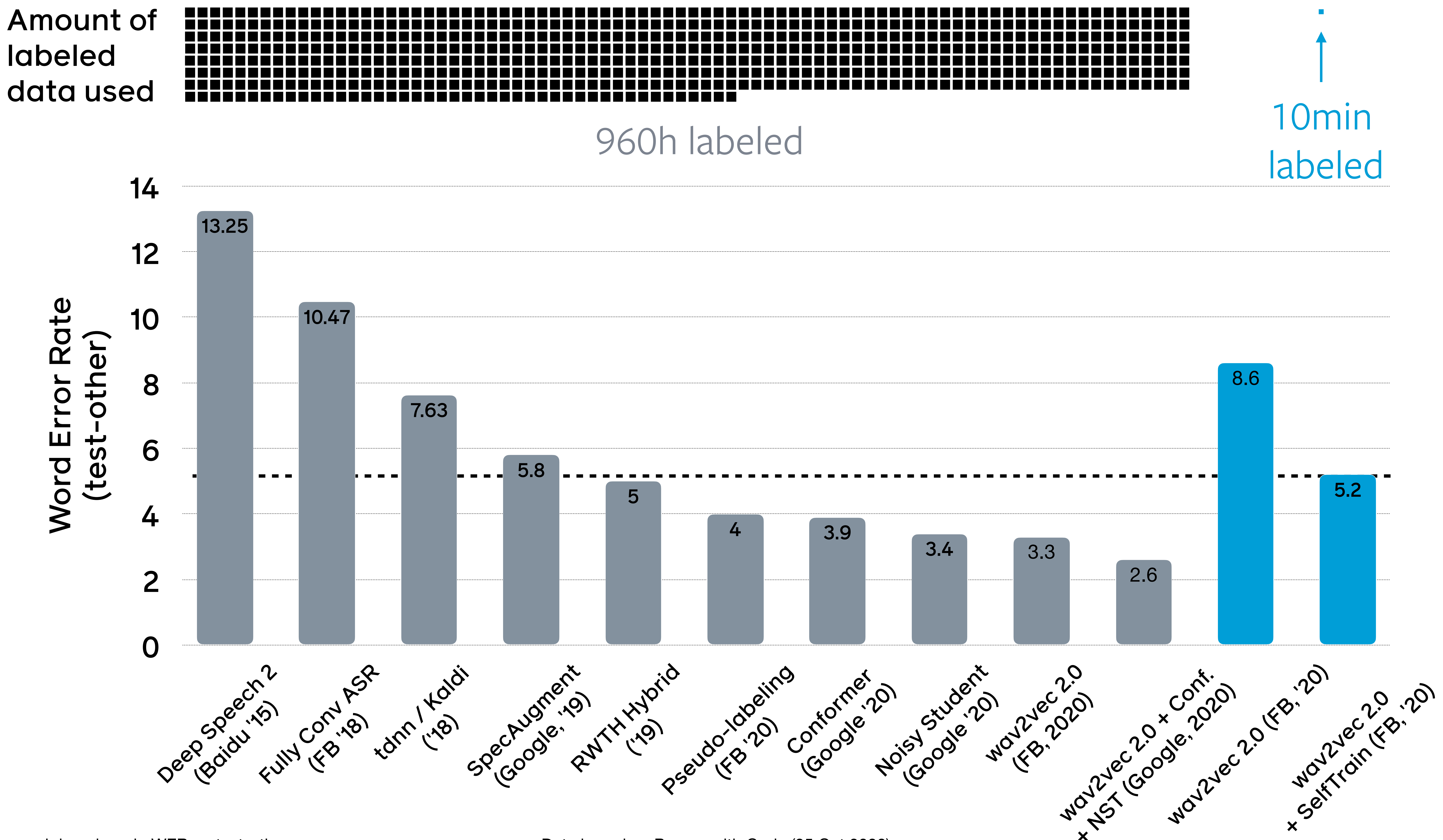
- Self-training very successful in speech recognition: generate pseudo-labels



Pre-training and self-training

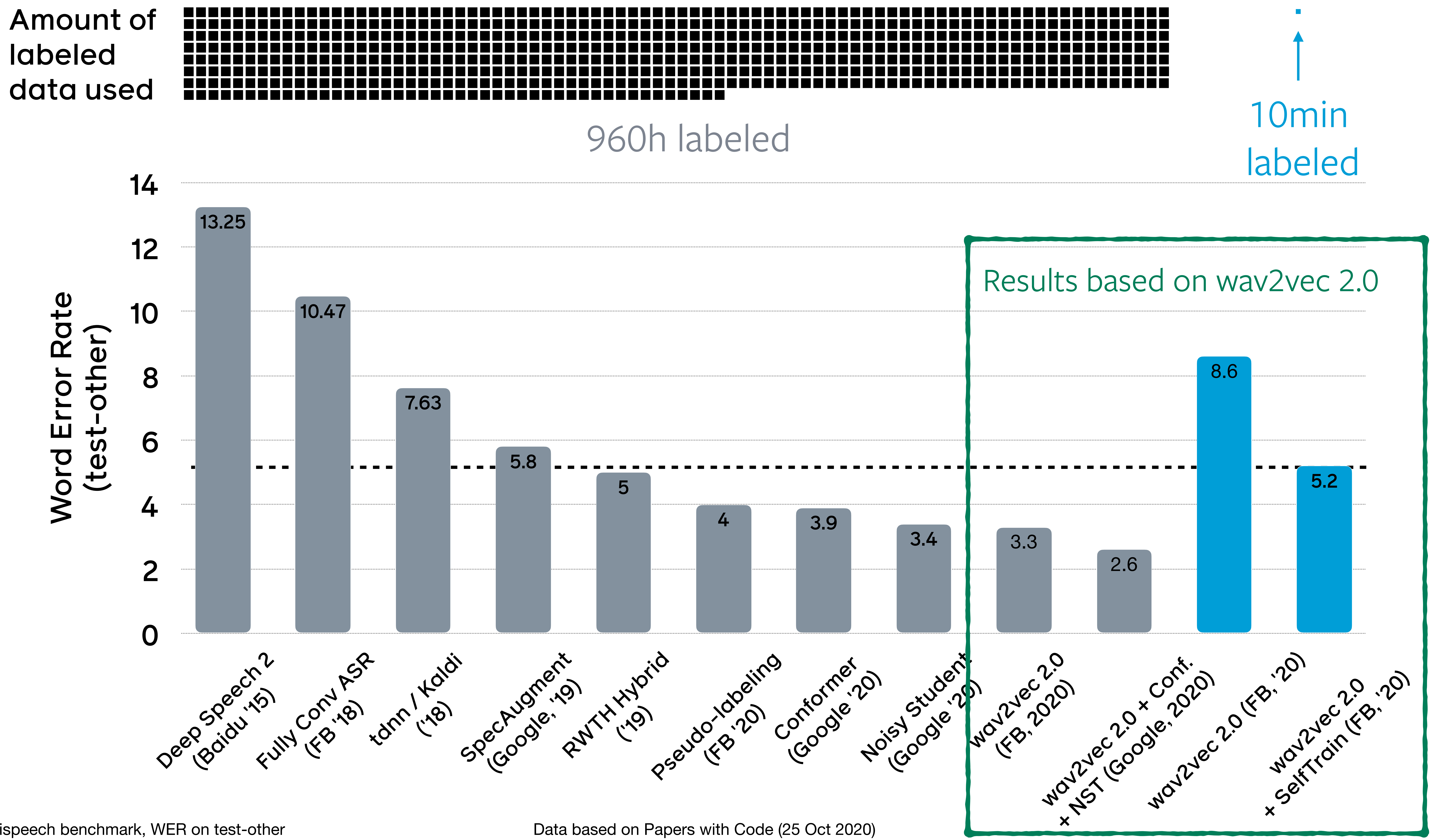
- Self-training very successful in speech recognition: generate pseudo-labels





Librispeech benchmark, WER on test-other

Data based on Papers with Code (25 Oct 2020)



Librispeech benchmark, WER on test-other

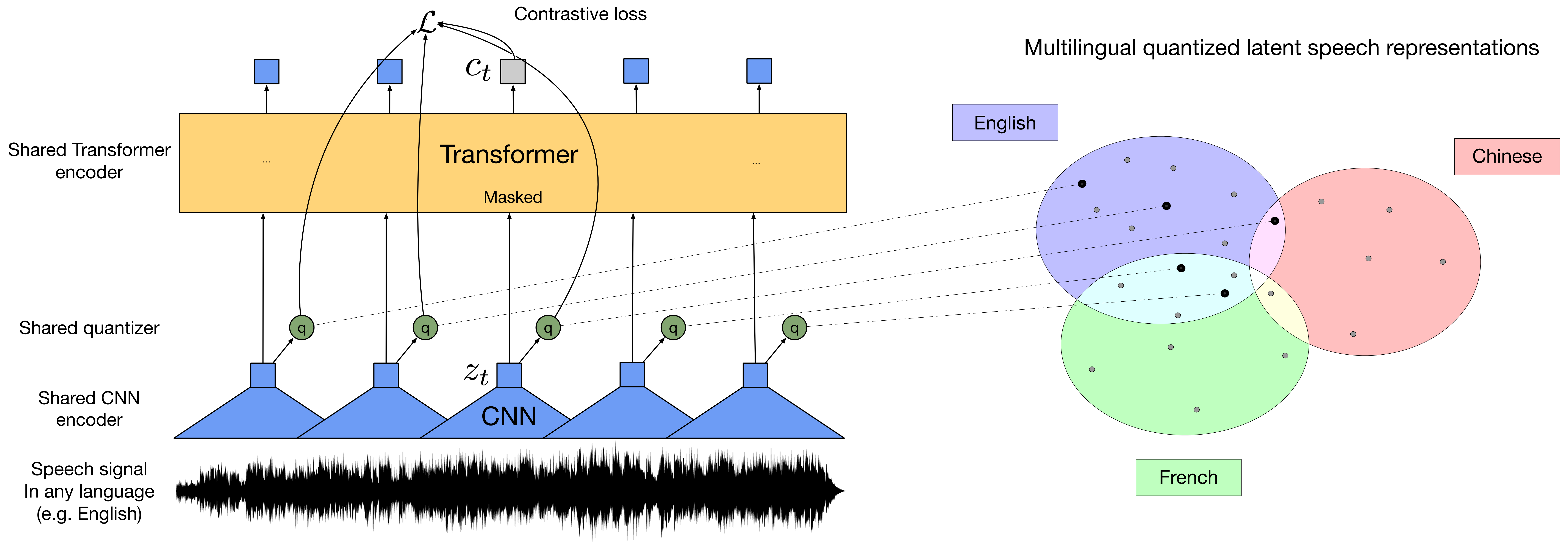
Data based on Papers with Code (25 Oct 2020)

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

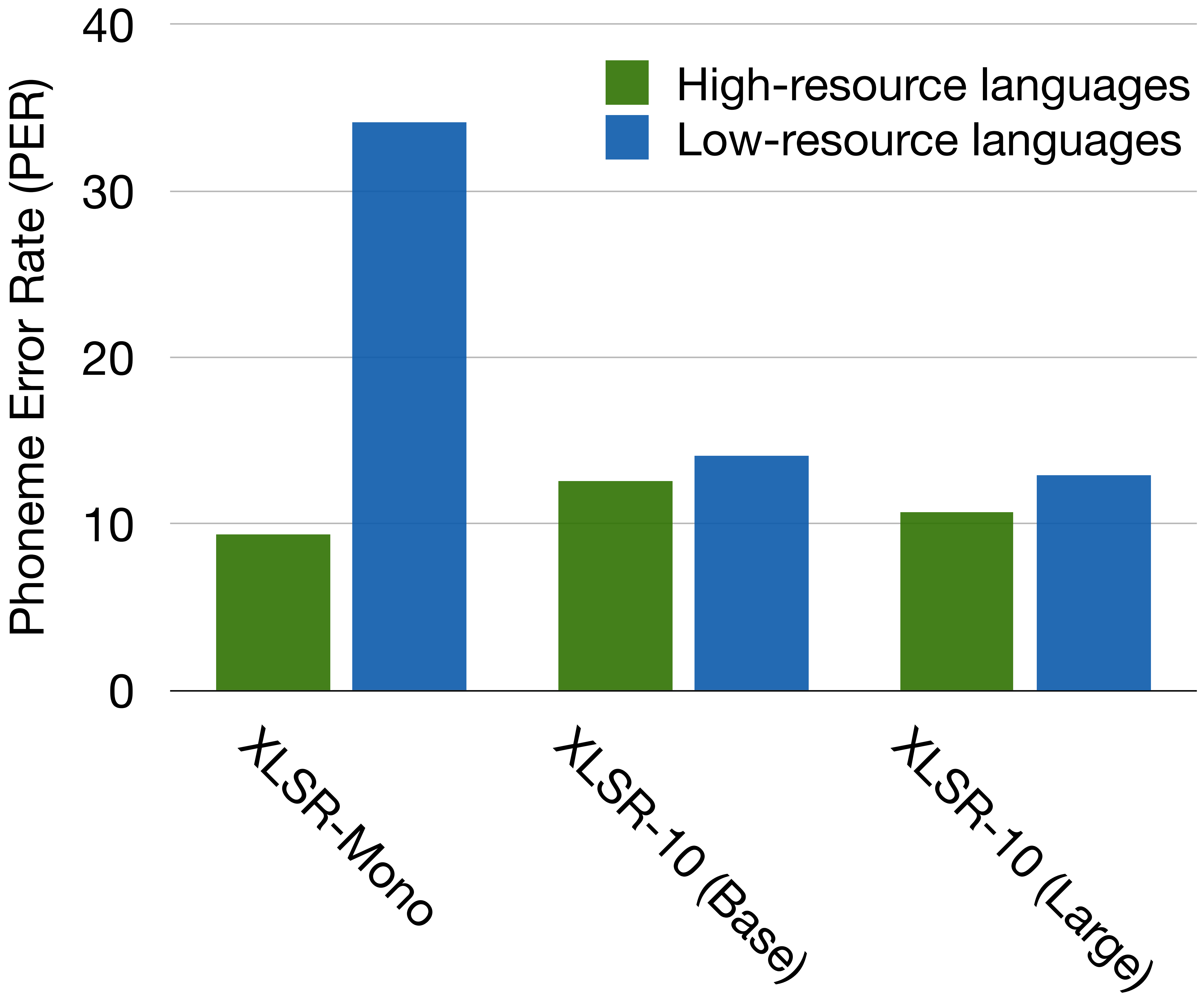
XLSR: cross lingual speech representation learning with wav2vec



XLSR: Results - cross-lingual transfer

Cross-lingual transfer = Train data from high-resource languages benefits low-resource languages.

CommonVoice results:



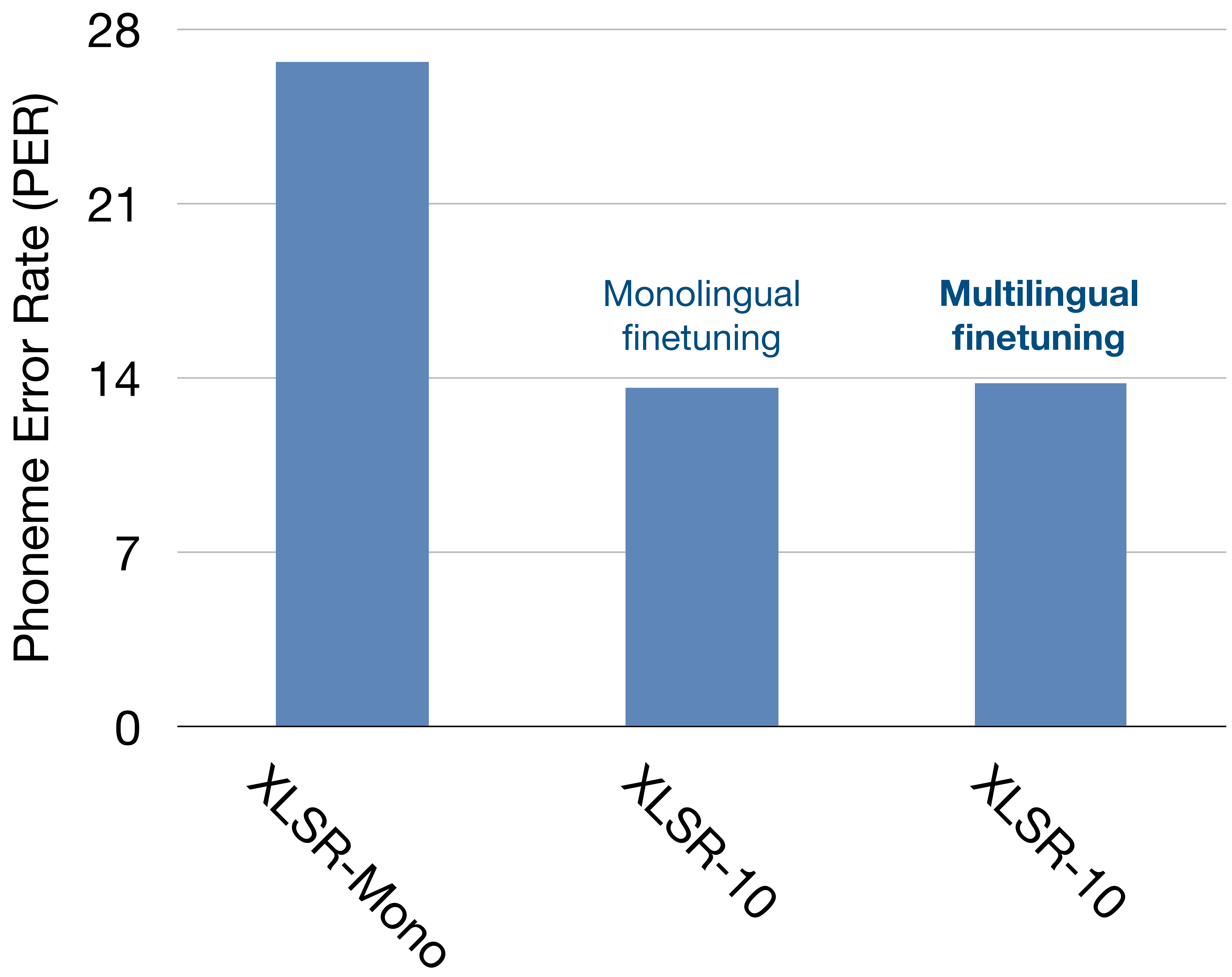
XLSR: Results - multilingual fine-tuning

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

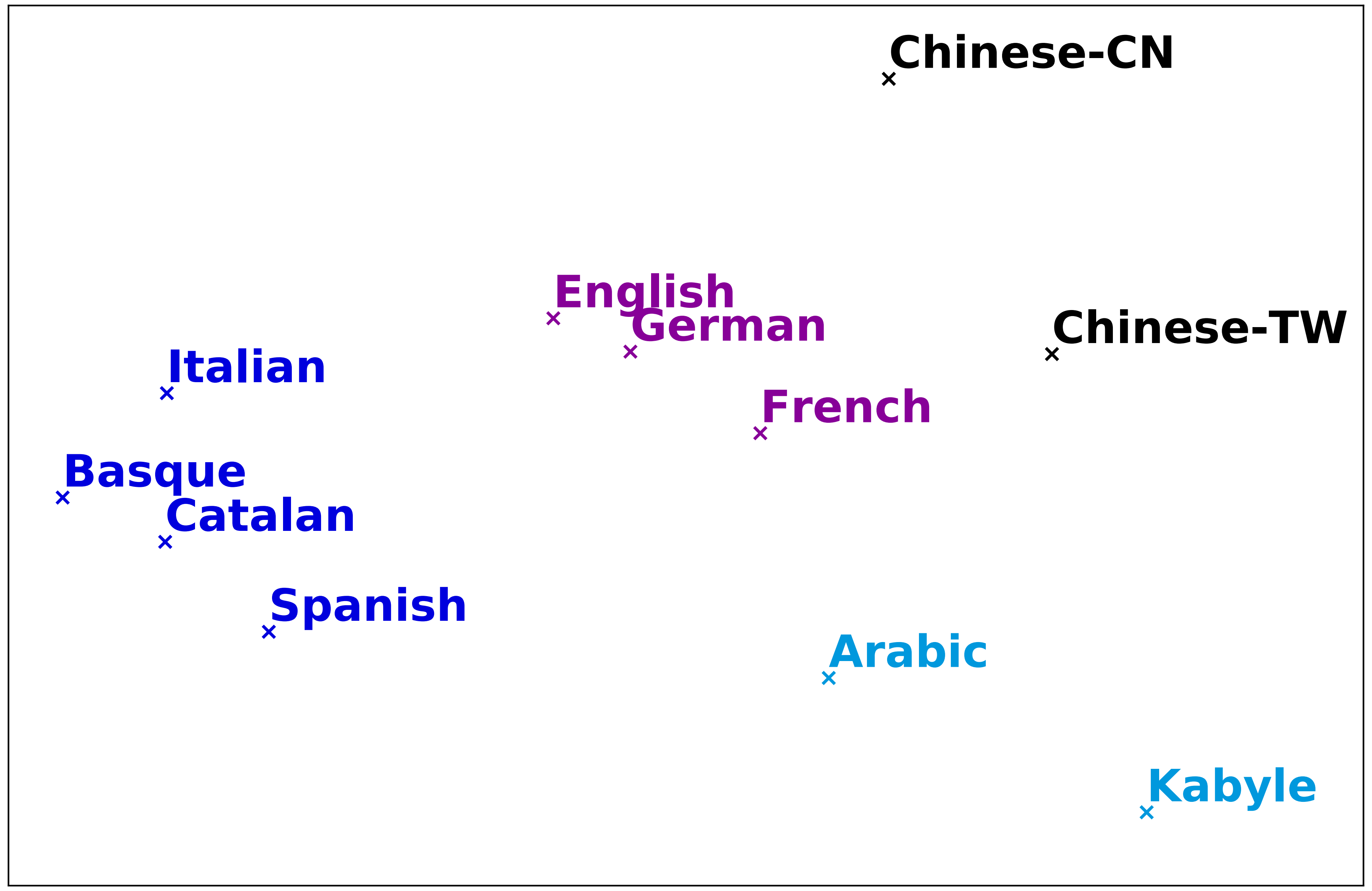
CommonVoice results:

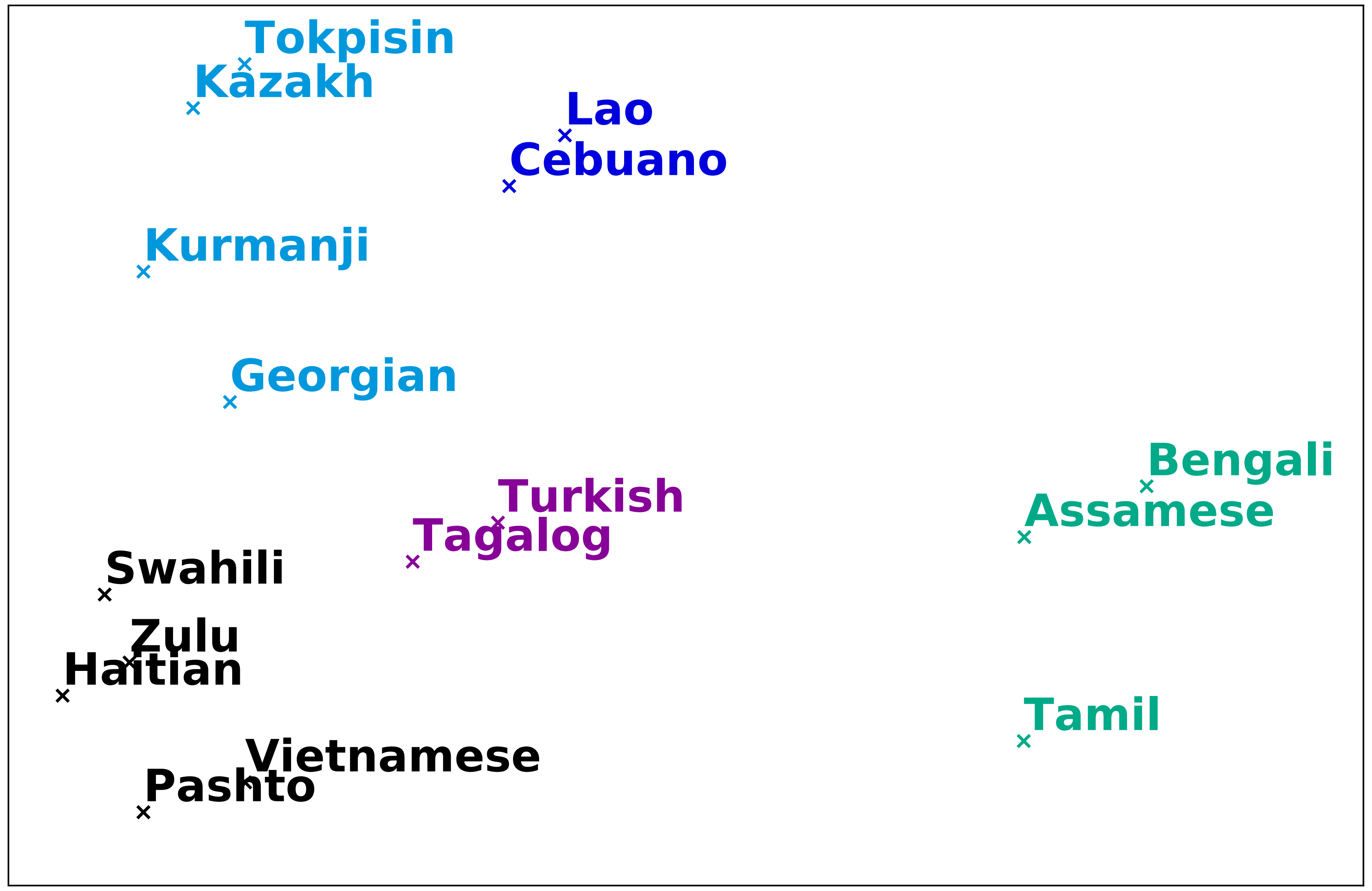


XLSR: Analysis of discrete latent speech representations

PCA visualization of latent discrete representations from the multilingual codebook

Similar languages tend to share discrete tokens and thus cluster together





Unsupervised Speech Recognition

Unsupervised speech recognition

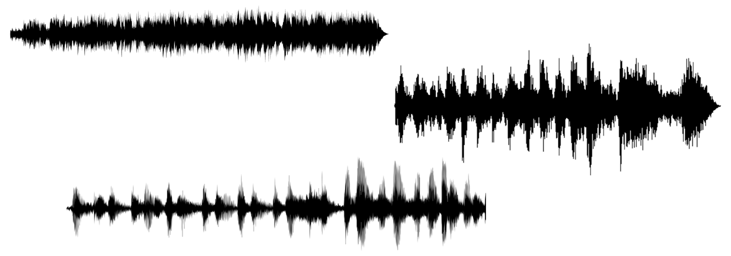
- Entirely remove need for labeled data
- Unsupervised machine translation works*, what about speech?
- Key problem: what are the units in the speech audio?

wav2vec Unsupervised: Key ideas

- Learn good representations of speech audio
- Unsupervised segmentation of the speech audio into phonemic units
- Learn mapping between speech segments and phonemes using adversarial learning

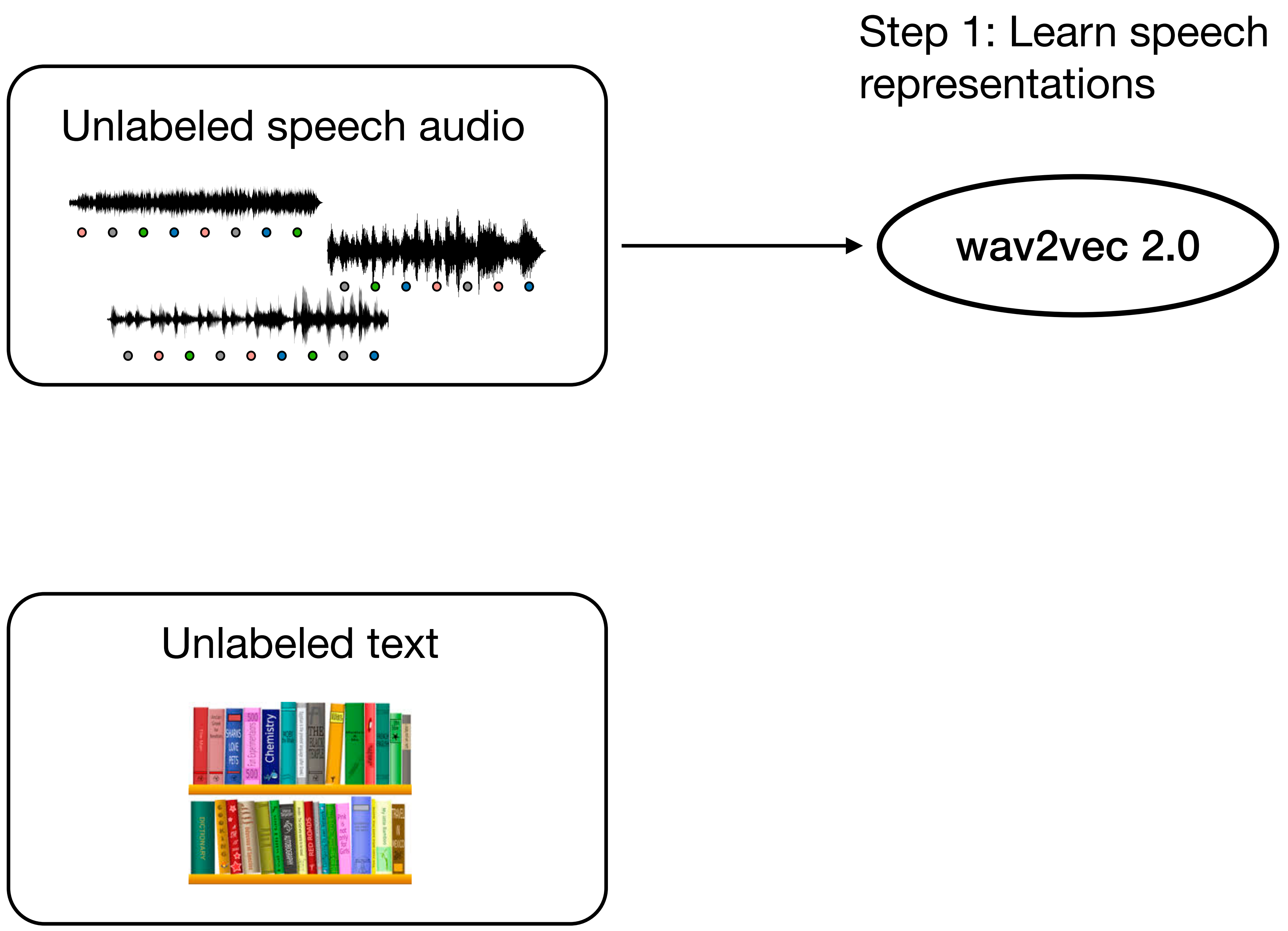
wav2vec Unsupervised

Unlabeled speech audio

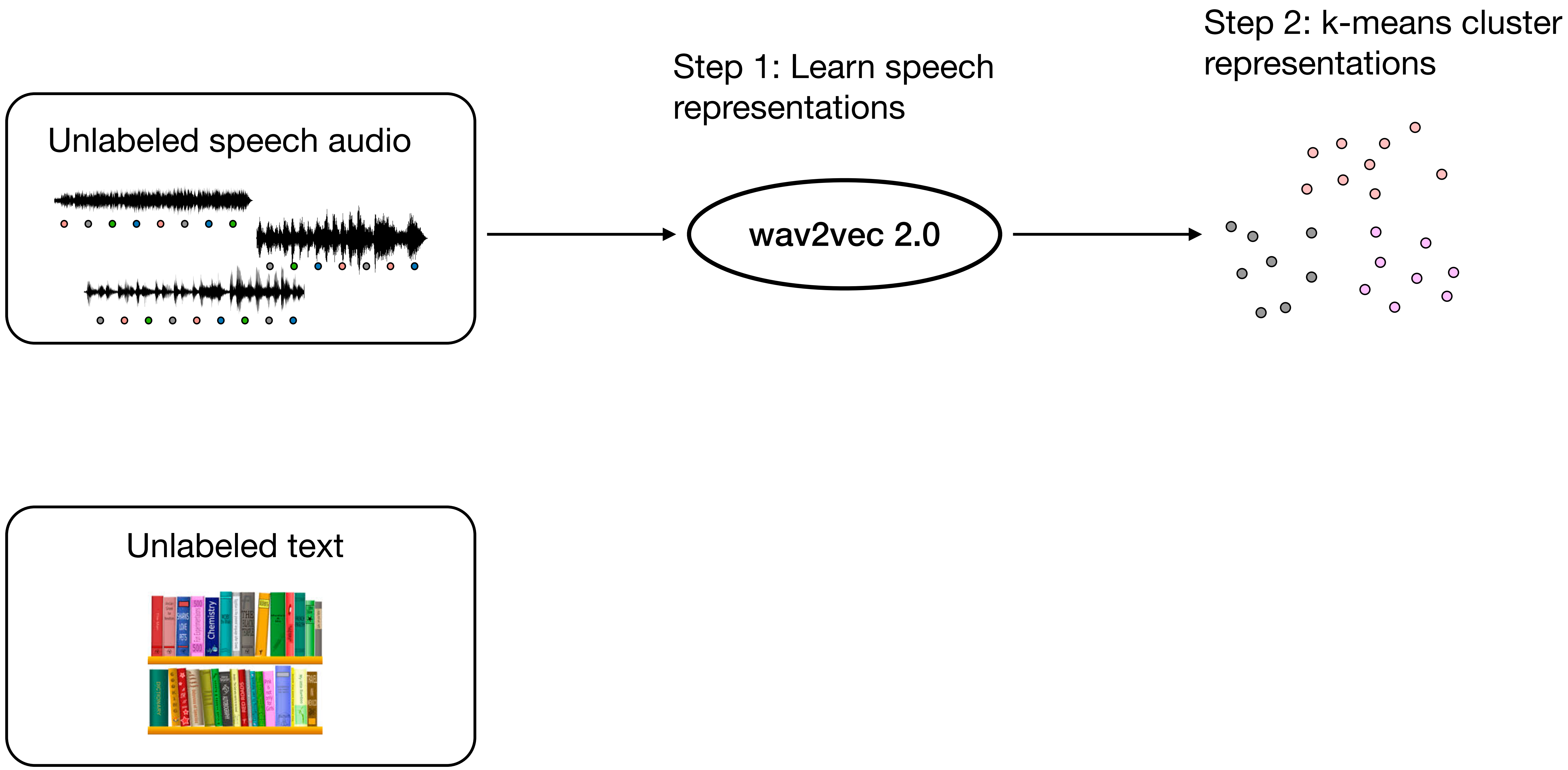


Unlabeled text

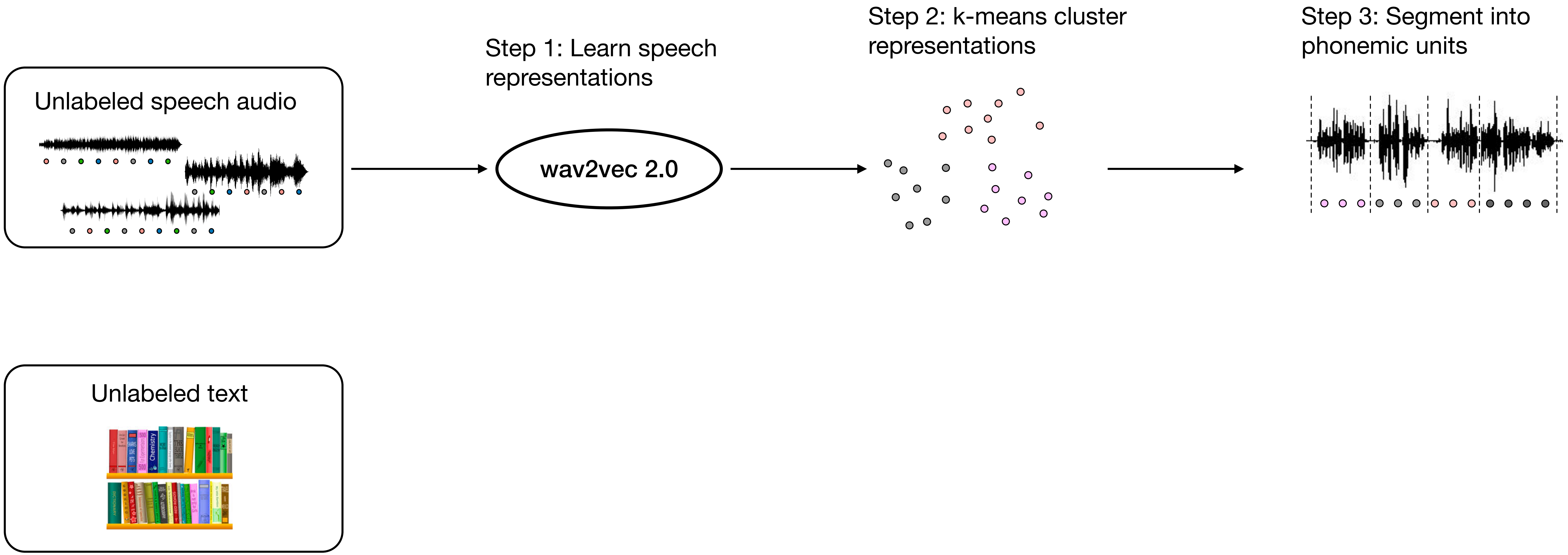
wav2vec Unsupervised



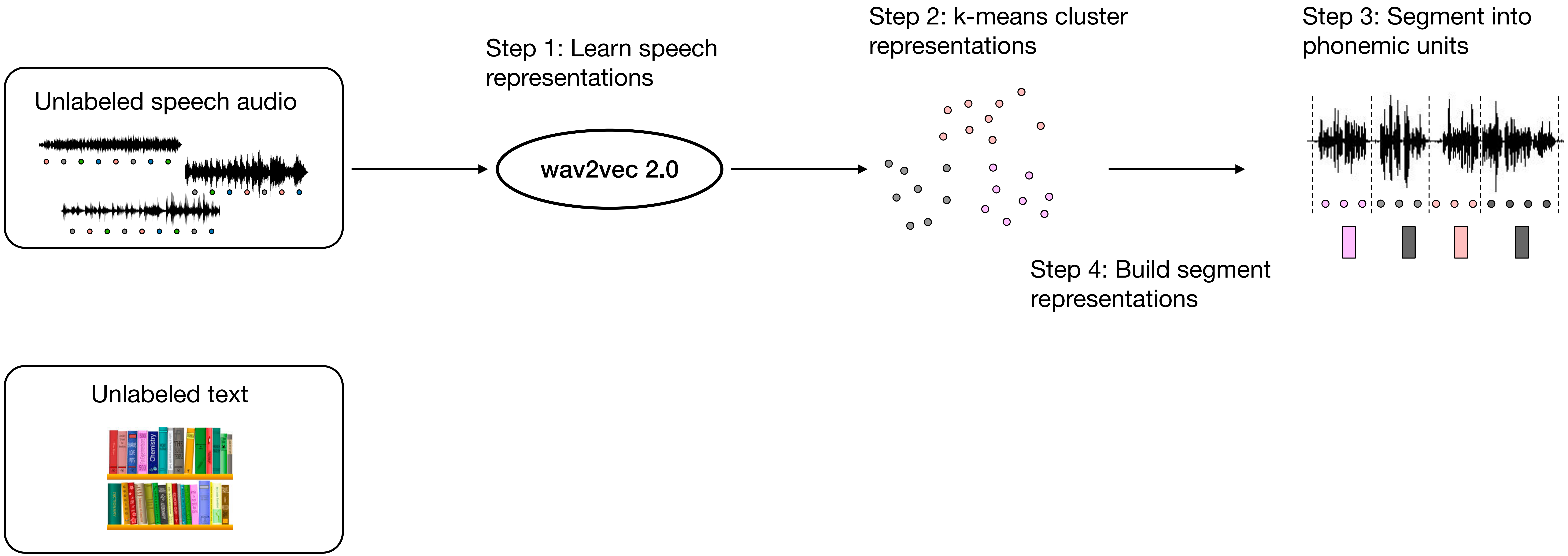
wav2vec Unsupervised



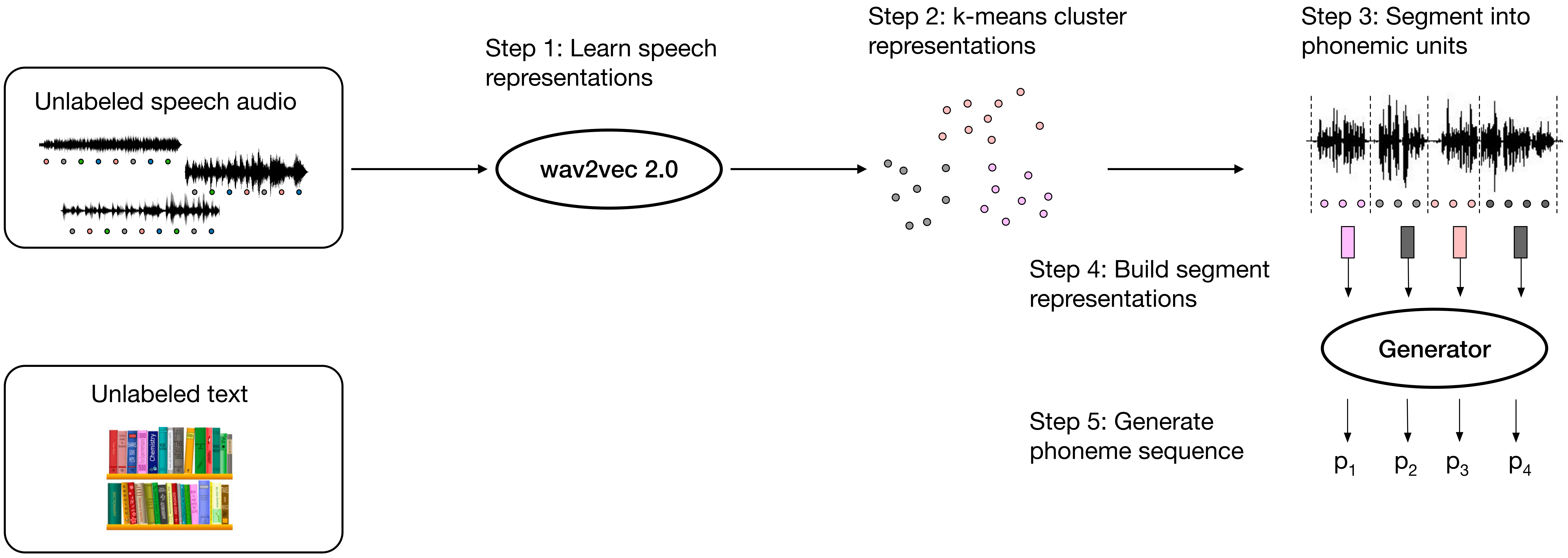
wav2vec Unsupervised



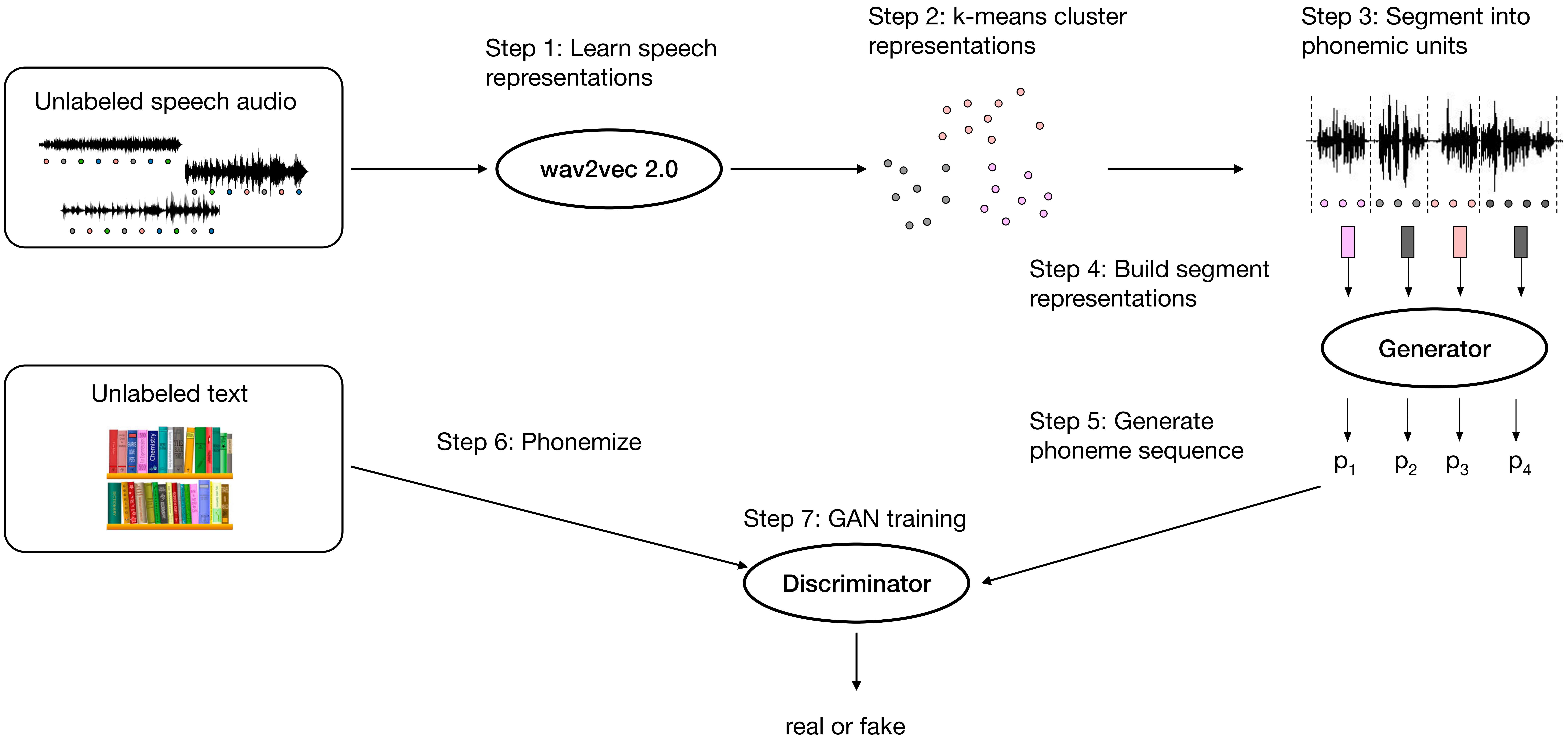
wav2vec Unsupervised



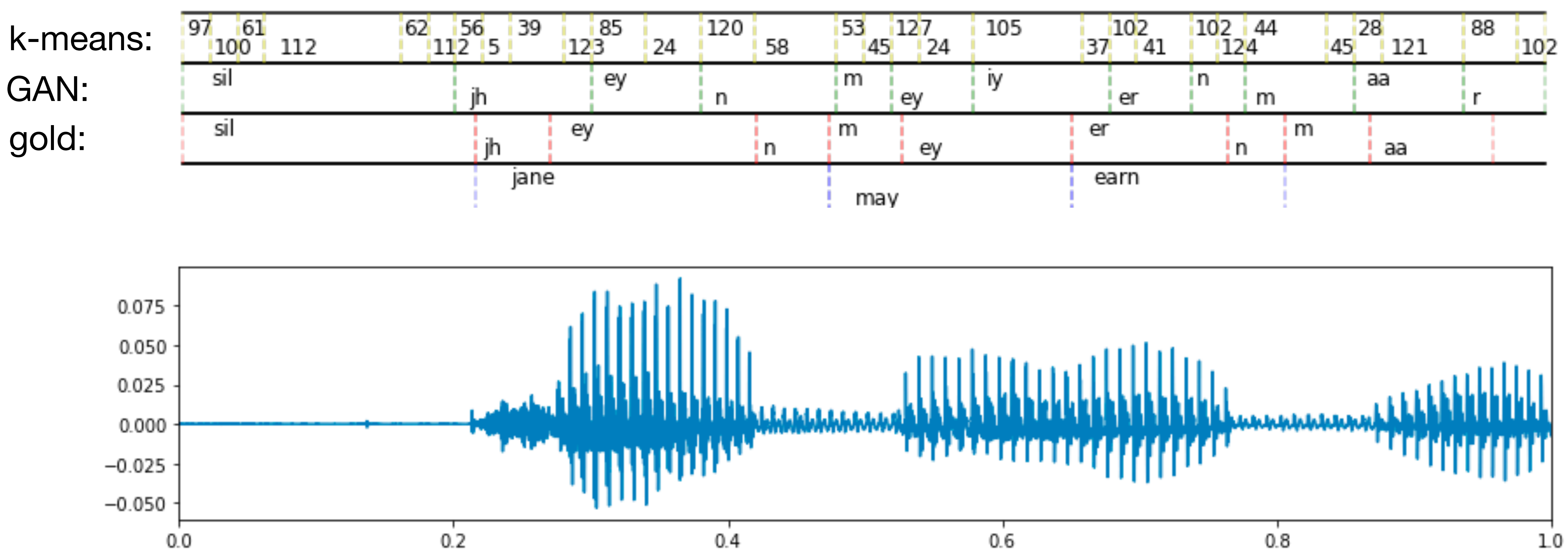
wav2vec Unsupervised



wav2vec Unsupervised



Simple segmentation



Text data pre-processing

Unlabeled text

he spoke soothingly

Text data pre-processing

Unlabeled text

he spoke soothingly

Phonemize

hh iy s ow k s uw dh ih ng l iy

Text data pre-processing

Unlabeled text

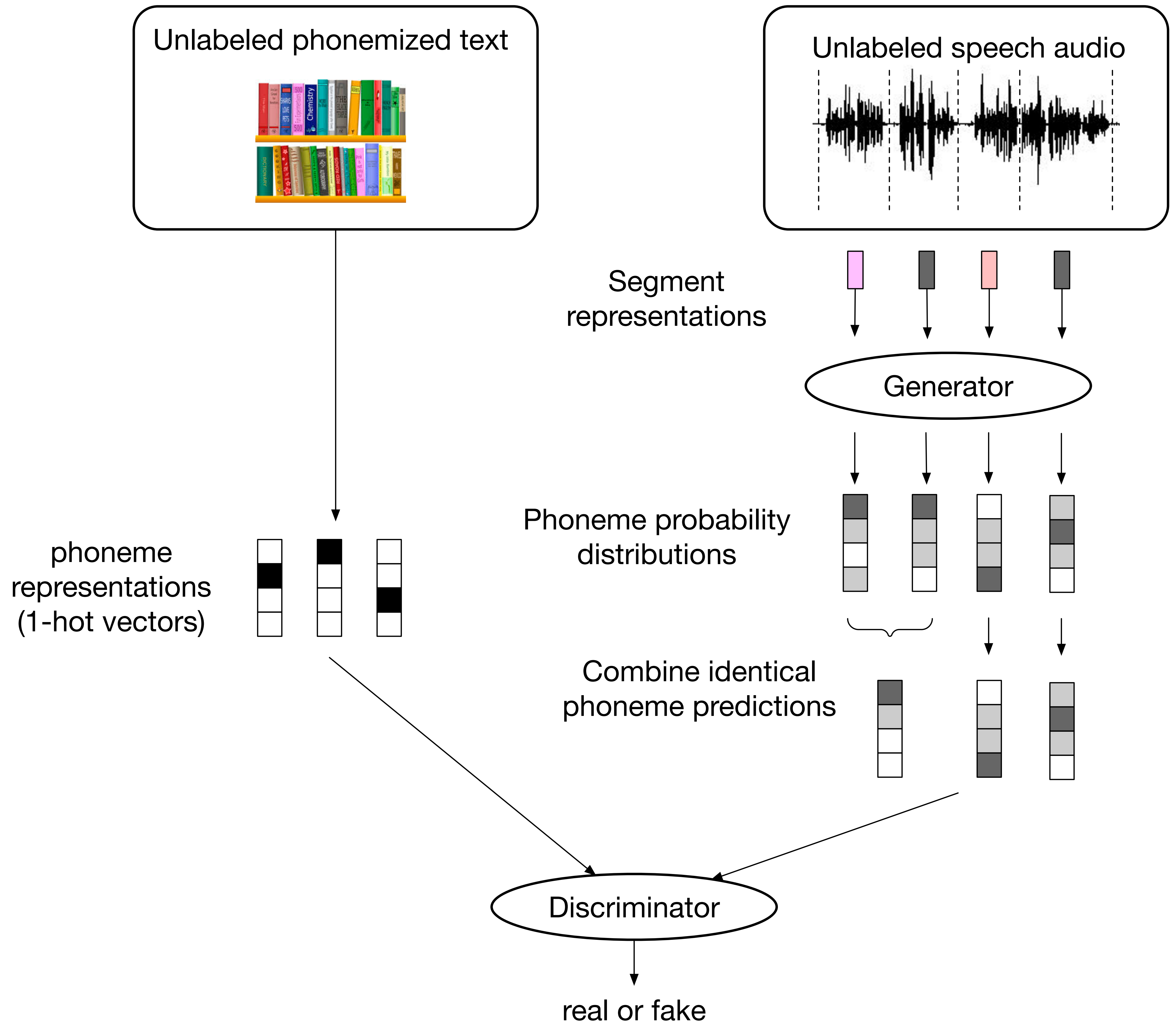
he spoke soothingly

Phonemize

sil hh iy s ow k s uw dh ih ng l iy sil

Silence insertion

GAN inputs

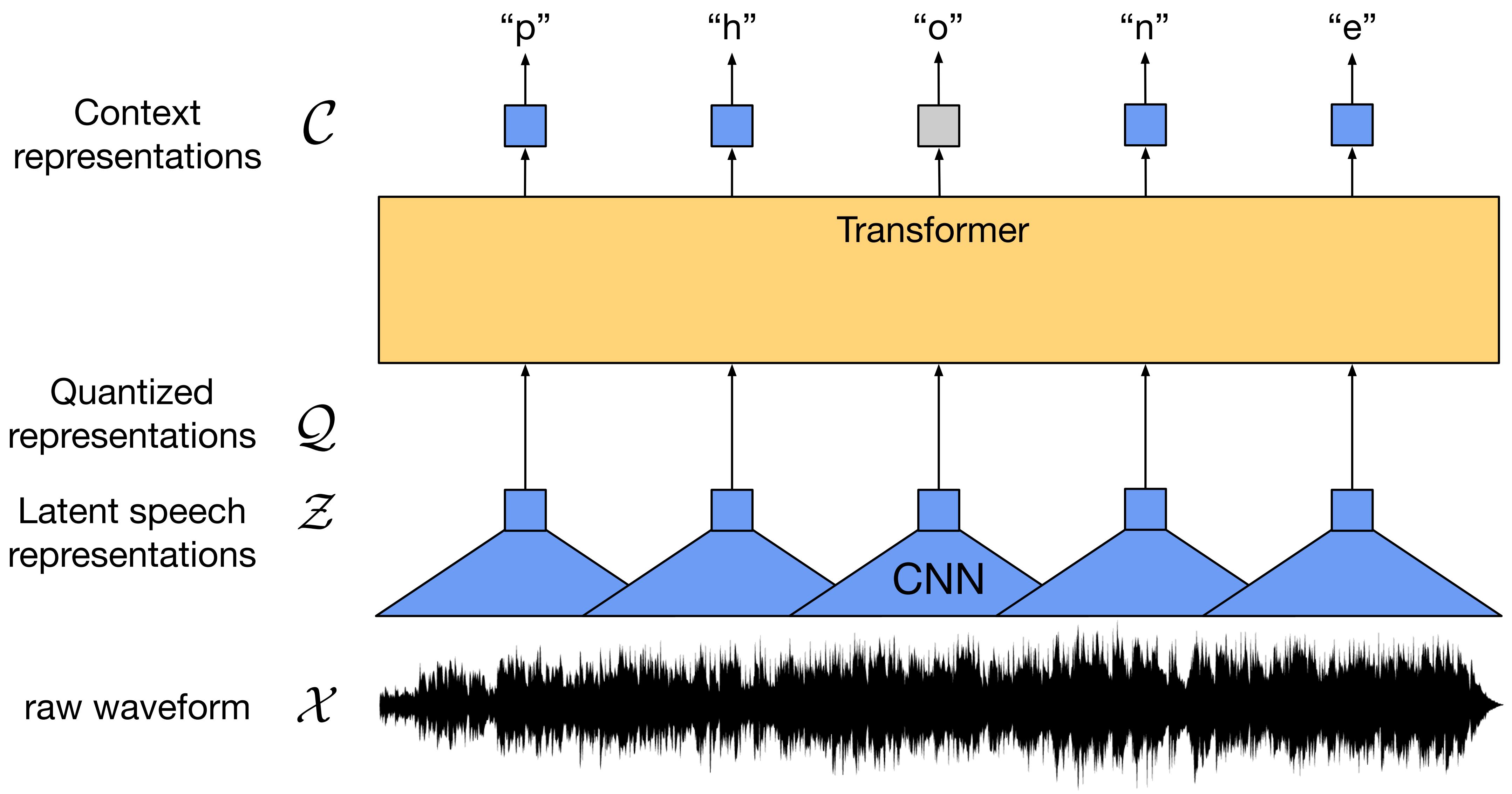

Unlabeled phonemized text
Unlabeled speech audio
Segment
representations
Generator
Phoneme probability
distributions
phoneme
representations
(1-hot vectors)
Combine identical
phoneme predictions
Discriminator
real or fake

Generator / Discriminator

- Generator: 1 layer CNN with 90k parameters
  w2v features frozen

- Discriminator: 3 layer CNN

- Train time: 12-15h on a single V100



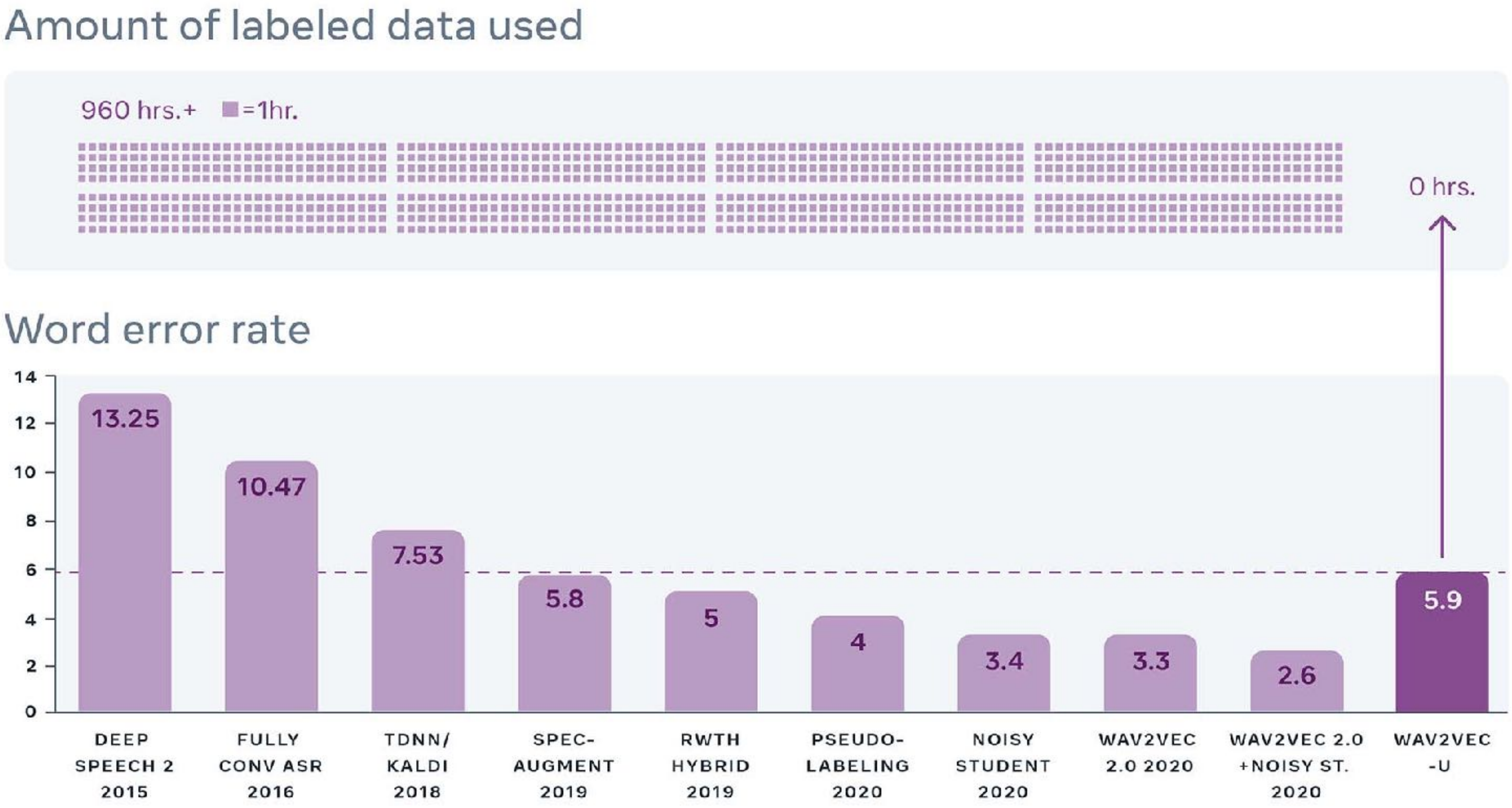
Training details

- Unsupervised metric for early stopping, hyper-parameter selection
- Self-training after GAN training (HMM and fine-tuning w2v)

Comparison to prior unsupervised work

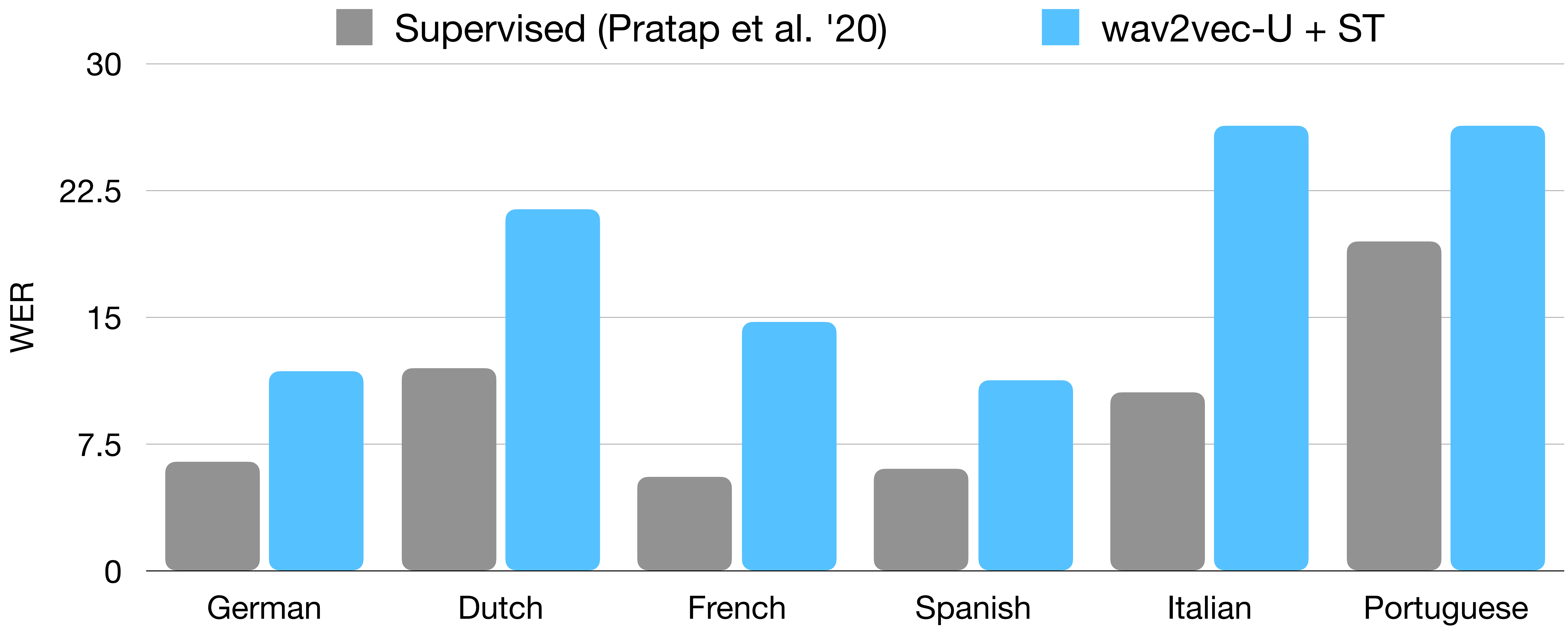
TIMIT Benchmark

Comparison to best supervised systems



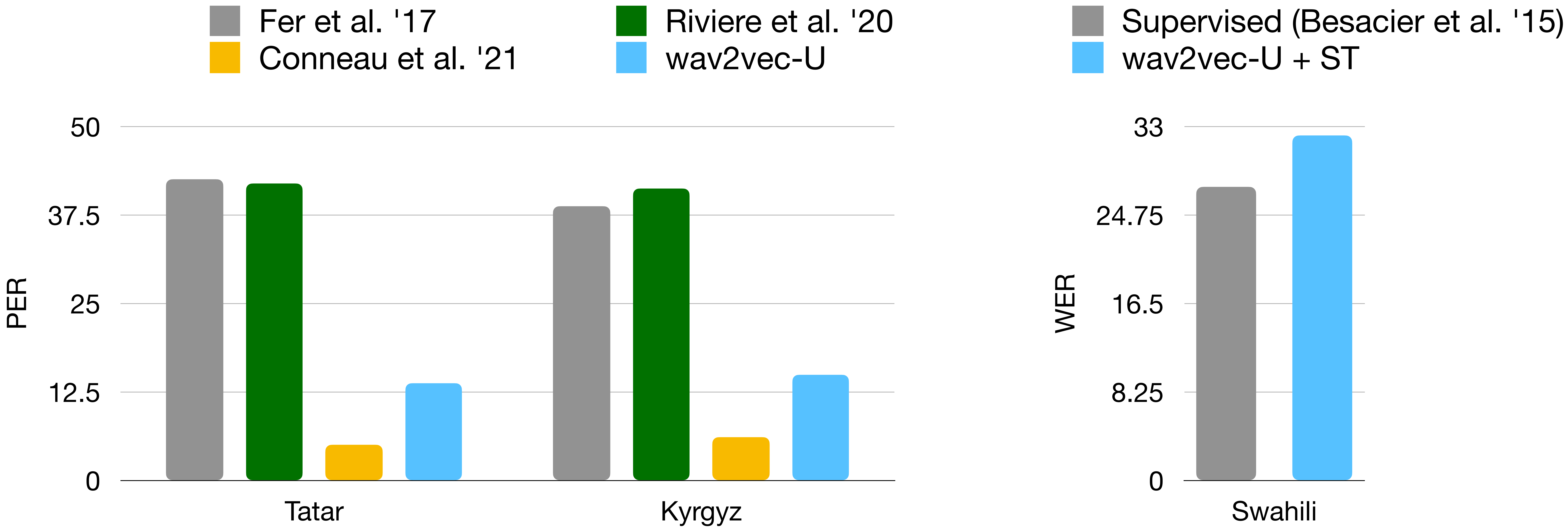
Librispeech benchmark

Other languages



MLS benchmark, wav2vec-U used only 100h of unlabeled data but there is up to 2k hours for some languages.

Low-resource languages



*wav2vec-U uses much less speech audio than prior work:
1.8h vs. 17h for Kyrgyz, 4.6h vs. 17h for Tatar

Discussion

- Very lightweight approach (except for wav2vec 2.0)
- Why does it work? Good audio features are main driver of performance
- Phonemizer still required
- Segment construction

Conclusion

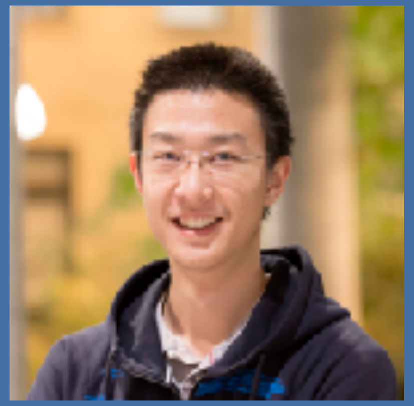
- Pre-training for speech works very well in both low-resource and high-resource setup.
- Cross-lingual training improves low-resource languages.
- Enable speech models with very little or even no labeled training data
- Make speech technology more ubiquitous and robust
- Code and models are available 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