CS11-737 Multilingual NLP

Speech Pre-training

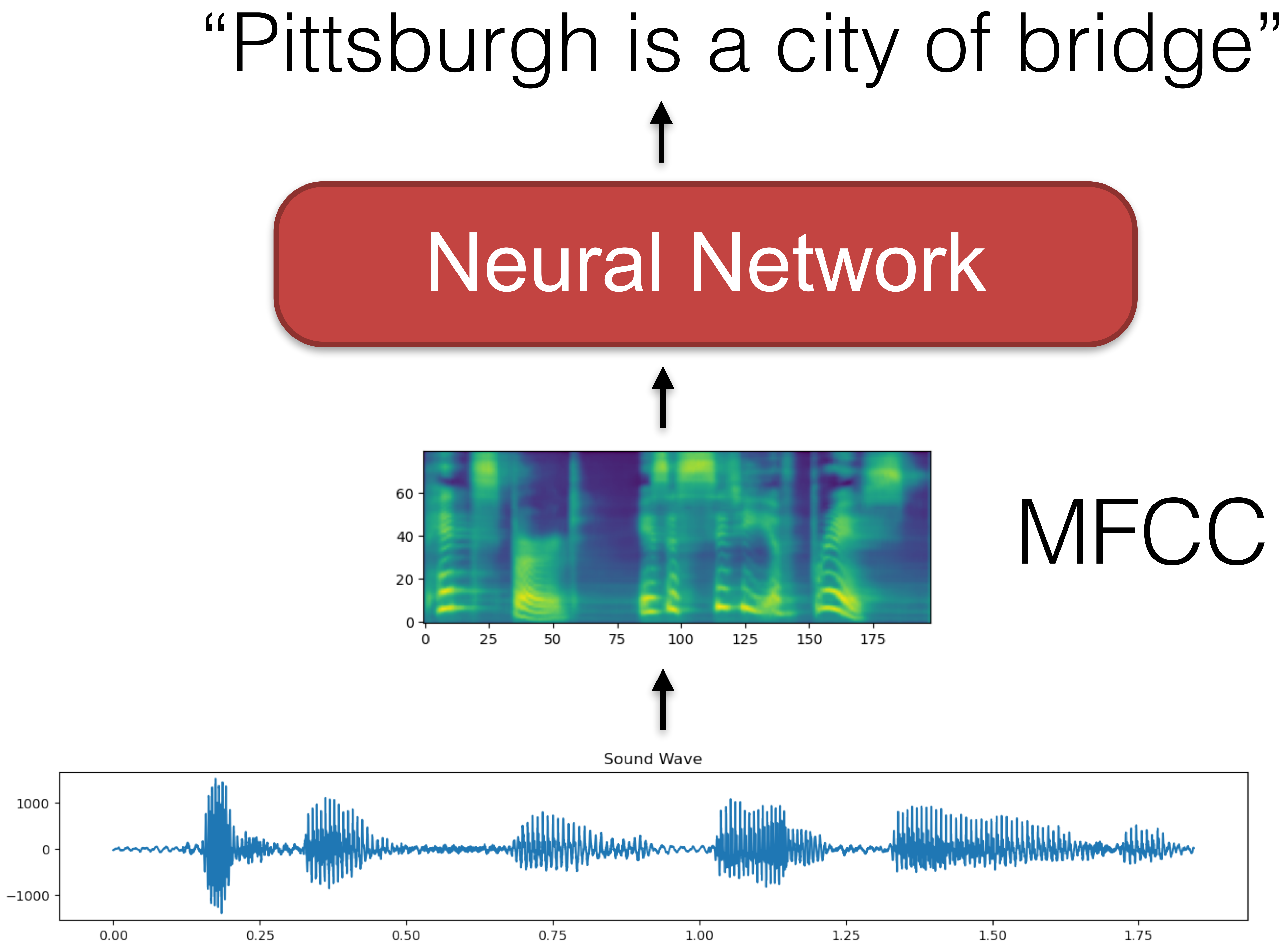
Lei Li

https://lileicc.github.io/course/11737mnlp23fa/

Carnegie Mellon University

Language Technologies Institute

Feature Extraction for Speech Recognition



“Pittsburgh is a city of bridge”

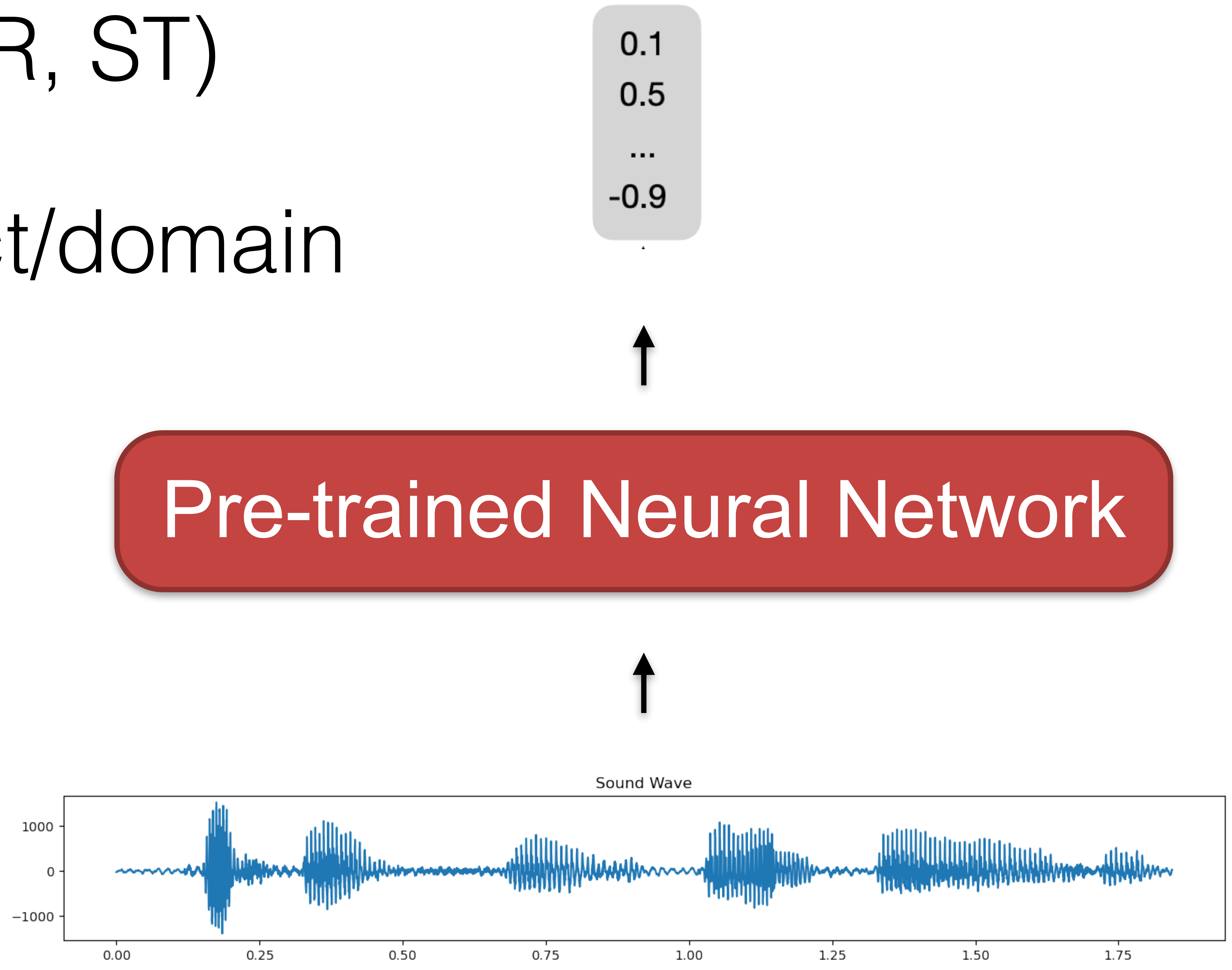
Neural Network

MFCC

- need 1,000+ hours of transcribed data to train a good ASR system
- how to generalize to many languages/dialects?

Self-supervised Speech Representation Learning

- Self-supervised Training on unlabeled audio data
- generalize to many tasks (ASR, ST)
- generalize to language/dialect/domain



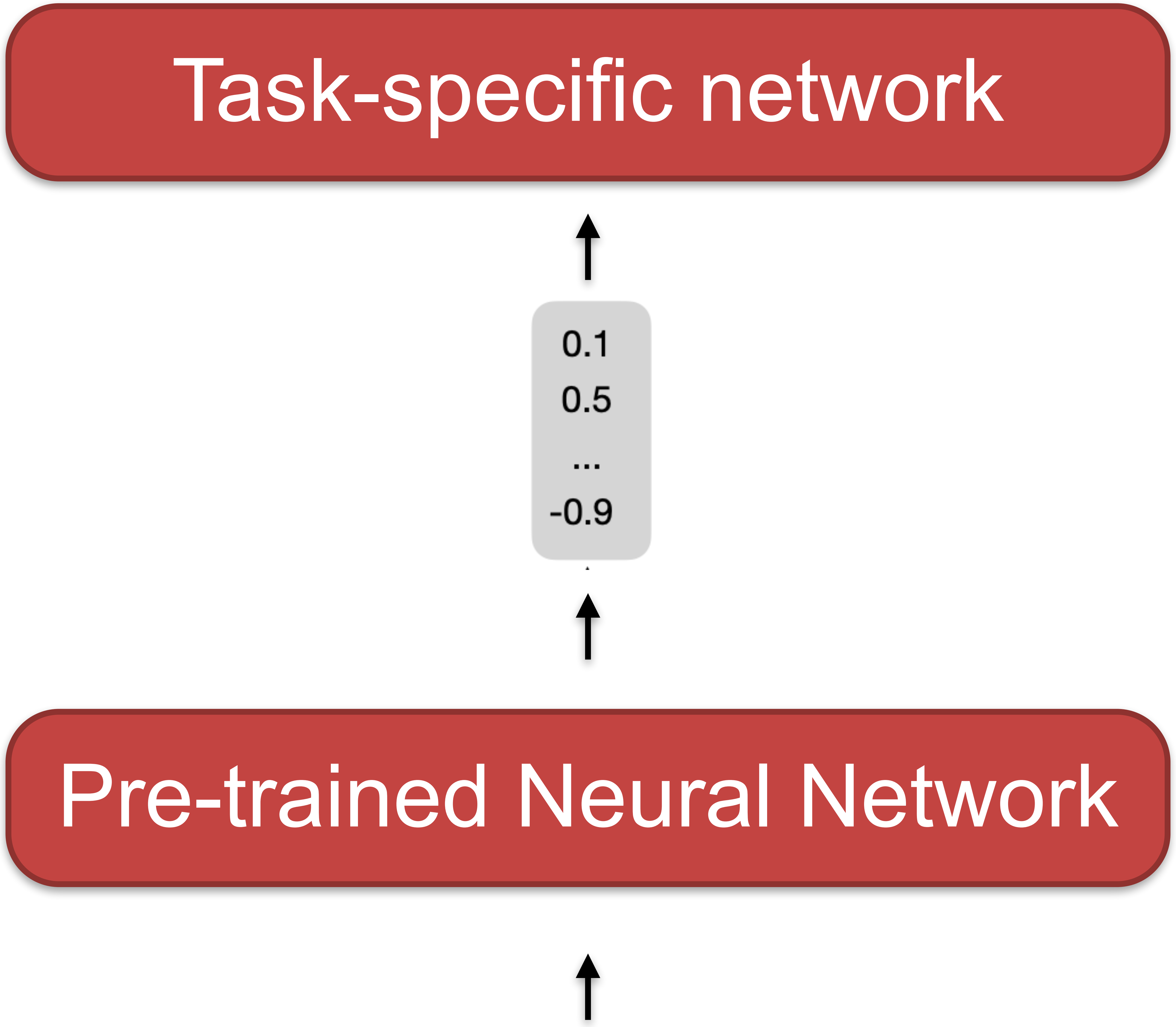
Pre-trained Neural Network

Transfer to Downstream Tasks

Fine-tuning

ASR

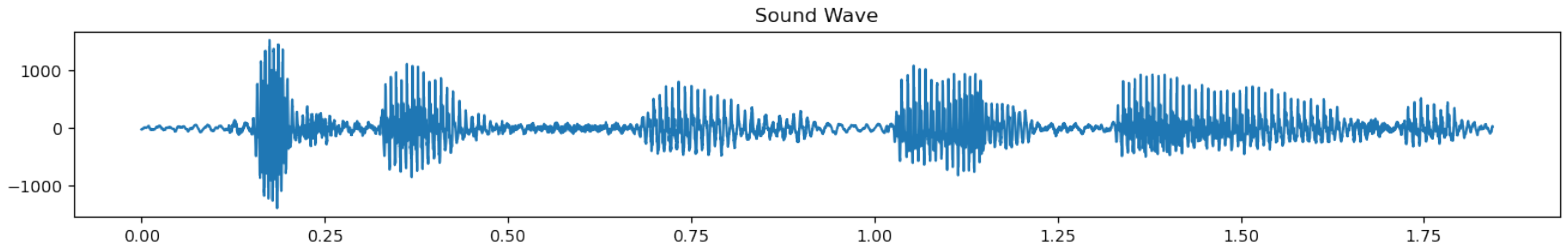
Speech Translation



Task-specific network

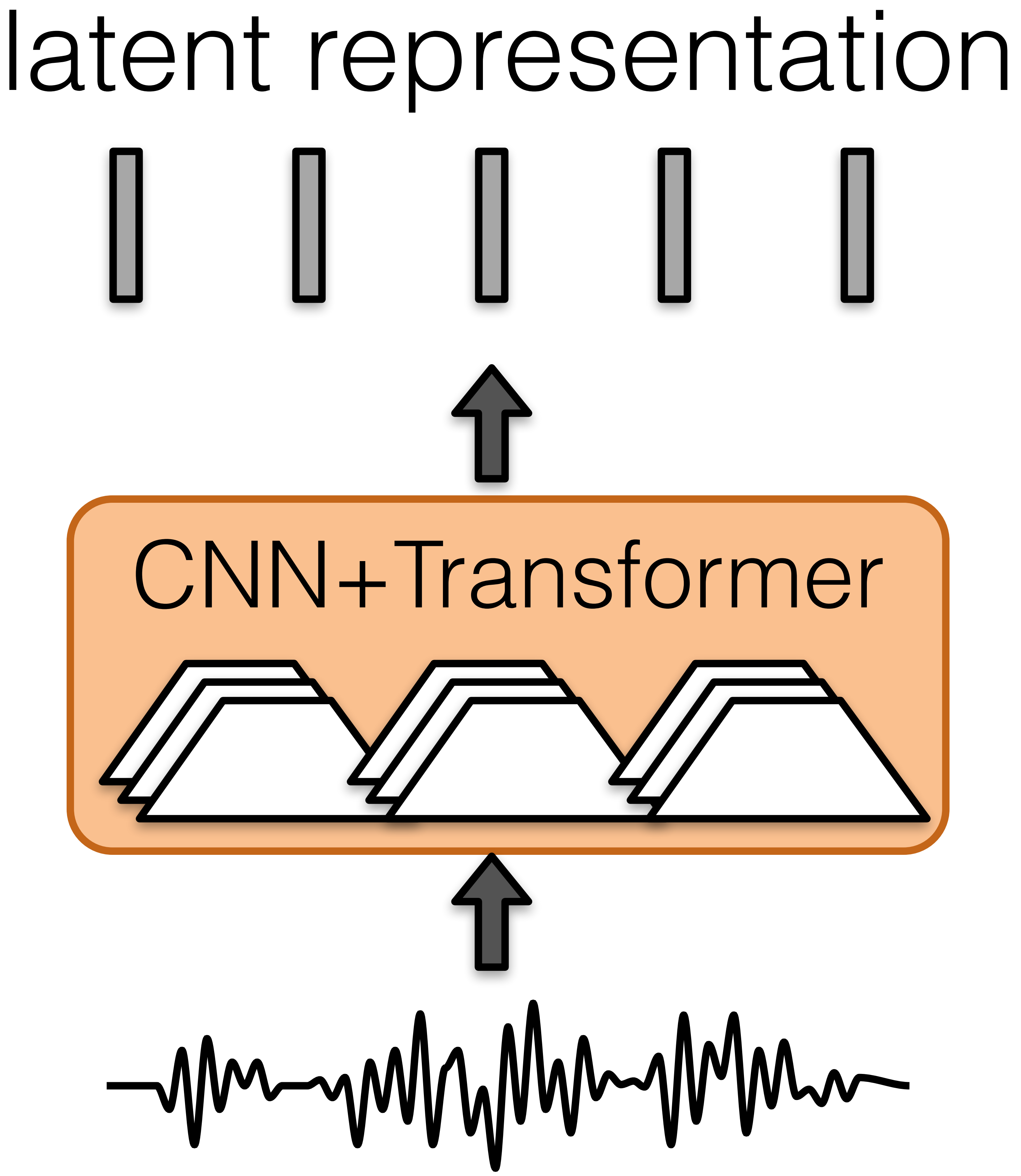
0.1
0.5
...
-0.9

Pre-trained Neural Network

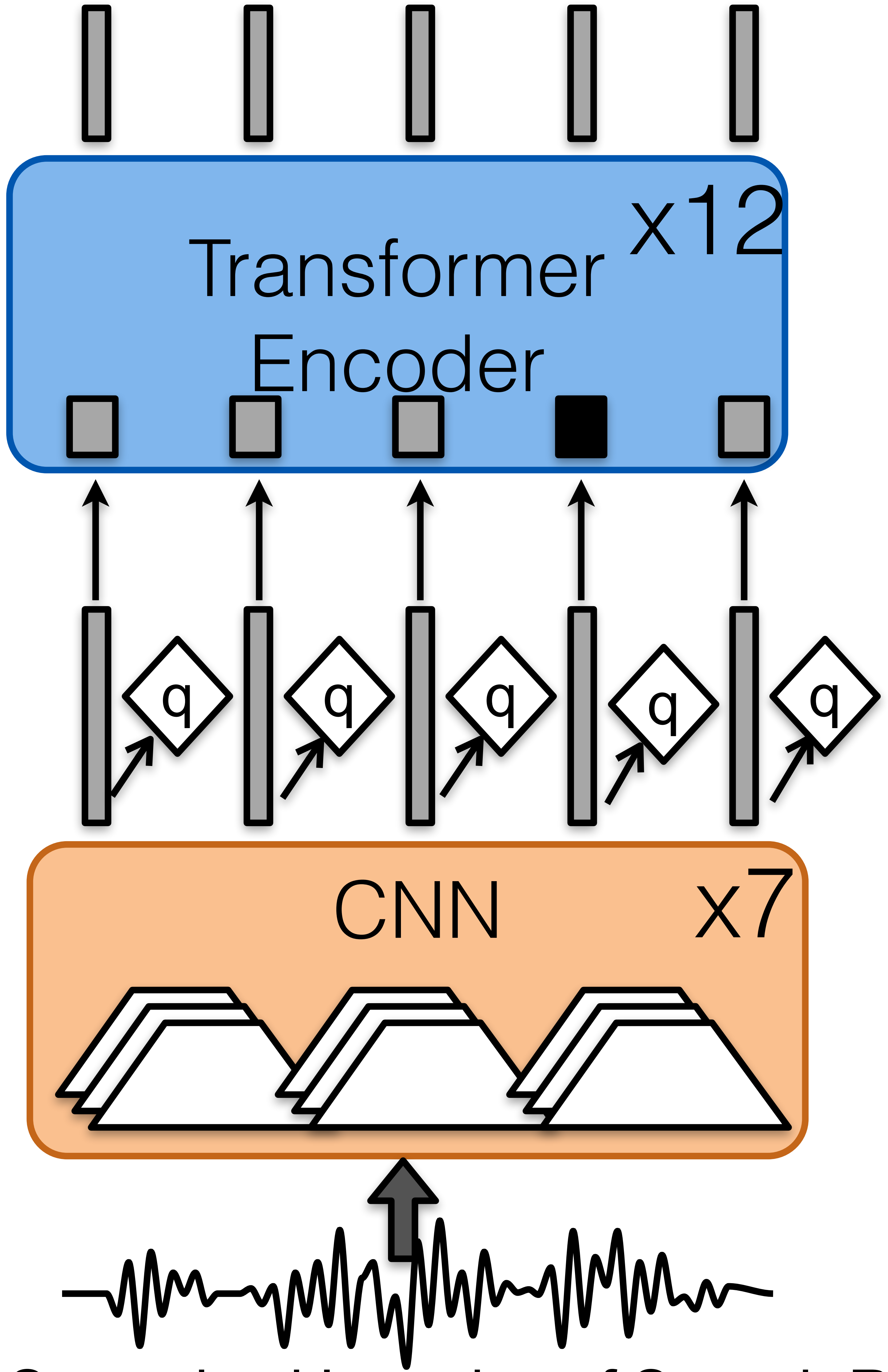


Wav2Vec / Wav2Vec 2.0

- Architecture:
  - CNN+Transformer
- Training
  - Masked prediction of quantized vector
  - contrast true quantized latent with distractor latent embeddings



Wav2Vec2



Context C

Mask during training

Quantized Rep Q

latent rep Z

Raw wav X, each frame ~ 25ms, stride 20ms

How many layers of Convolution?

How to design each kernel size/stride?

Wav2vec2.0: a Framework for Self-Supervised Learning of Speech Representations [Baevski et al, NeurIPS 2020]

Wav2Vec2

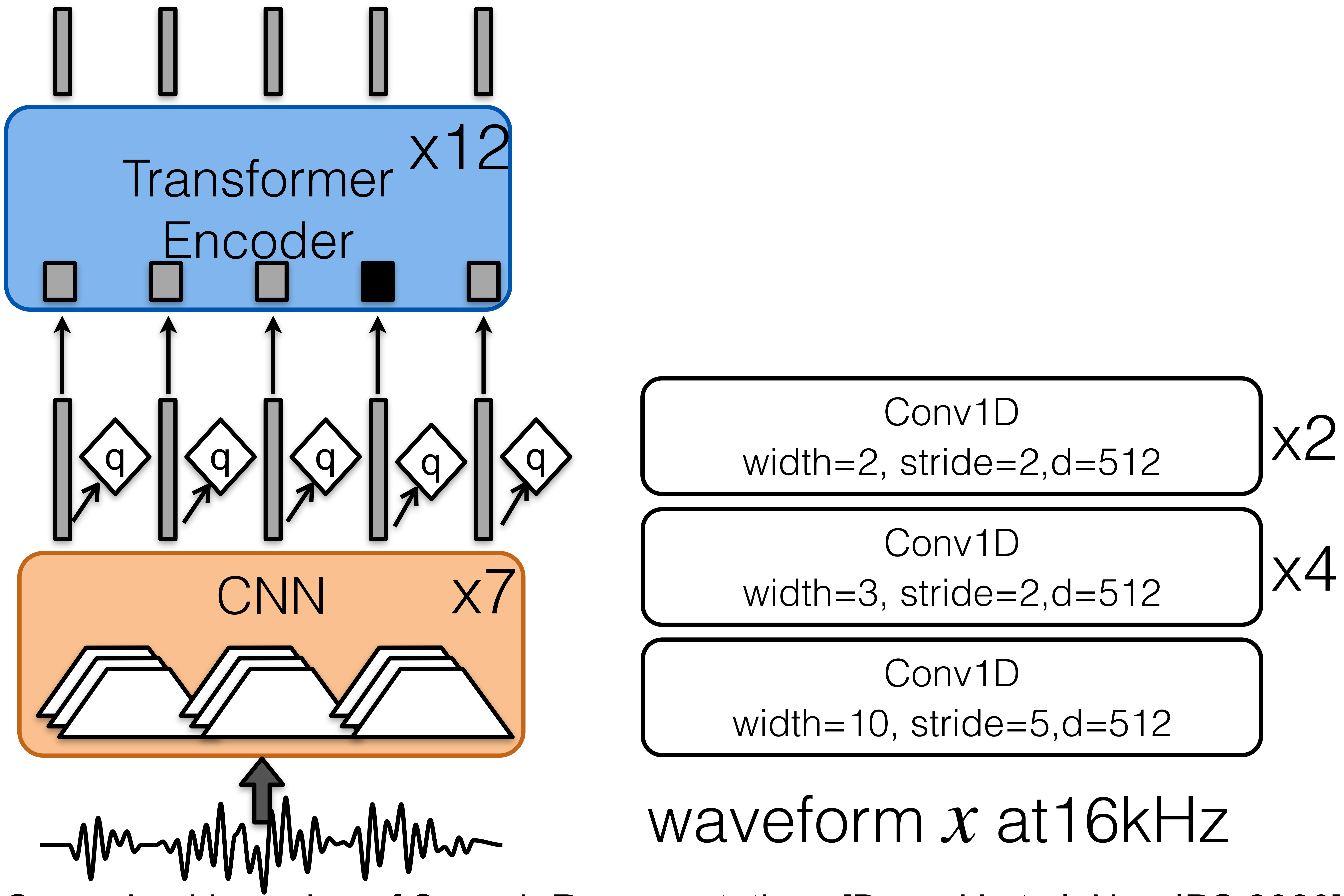
Context C

Mask during training

Quantized Rep Q

latent rep Z

Raw wav X, each frame ~ 25ms, stride 20ms



Transformer Encoder x12

CNN x7

Conv1D
width=2, stride=2,d=512 x2

Conv1D
width=3, stride=2,d=512 x4

Conv1D
width=10, stride=5,d=512

waveform x at16kHz

Wav2vec2.0: a Framework for Self-Supervised Learning of Speech Representations [Baevski et al, NeurIPS 2020]

Wav2Vec2

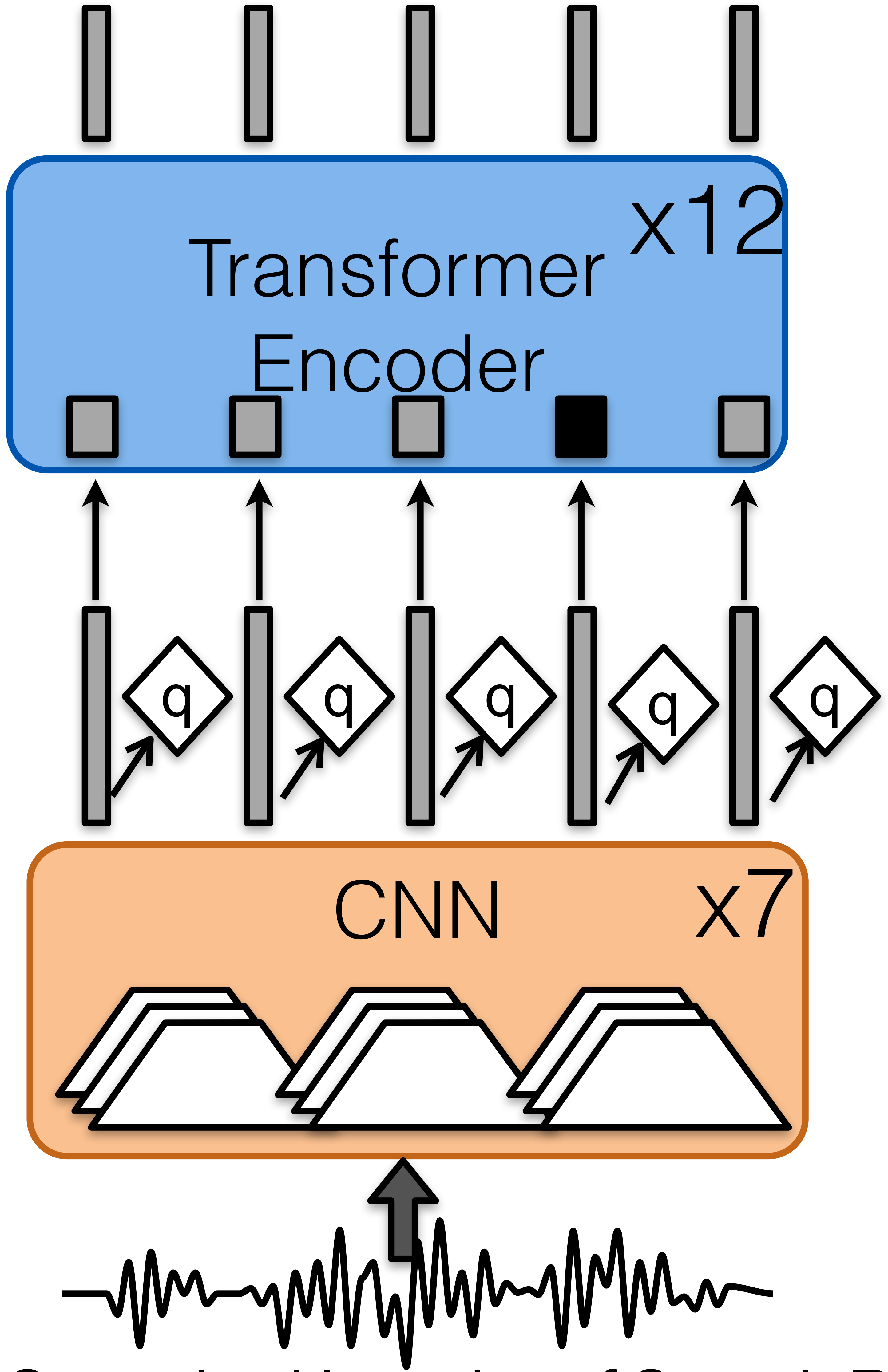
Context C

Mask during training

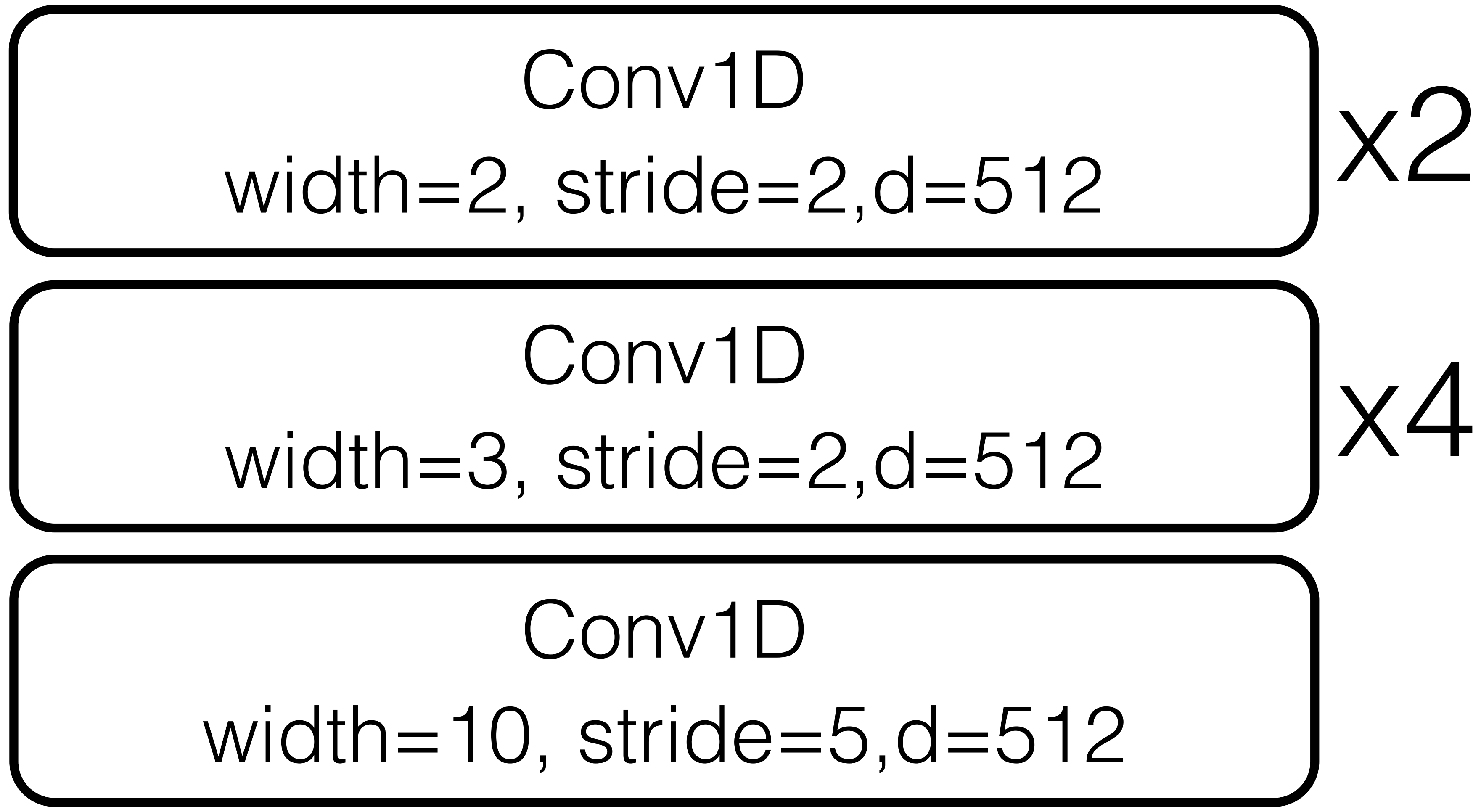
Quantized Rep Q

latent rep Z

Raw wav X, each frame ~ 25ms, stride 20ms



frame size=399 (25ms)
sampling rate=50Hz
(sliding 320=20ms)



Conv1D
width=2, stride=2,d=512
x2

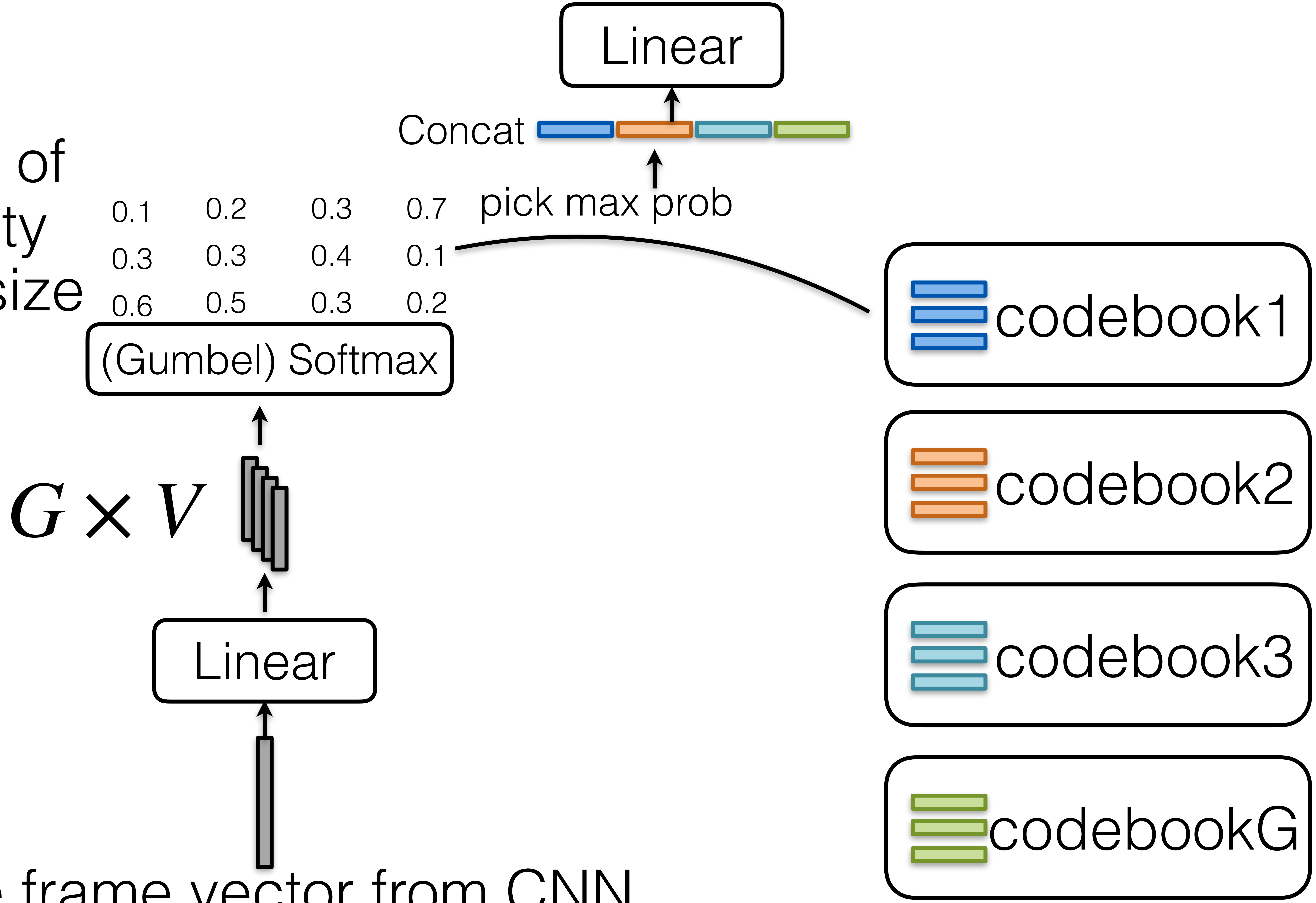
Conv1D
width=3, stride=2,d=512
x4

Conv1D
width=10, stride=5,d=512

waveform x at16kHz

Wav2vec2.0: a Framework for Self-Supervised Learning of Speech Representations [Baevski et al, NeurIPS 2020]

Discrete Quantization with Codebook



G groups of probability vector of size V

0.1	0.2	0.3	0.7
0.3	0.3	0.4	0.1
0.6	0.5	0.3	0.2

Linear

Concat

pick max prob

(Gumbel) Softmax

$G \times V$

Linear

codebook1

codebook2

codebook3

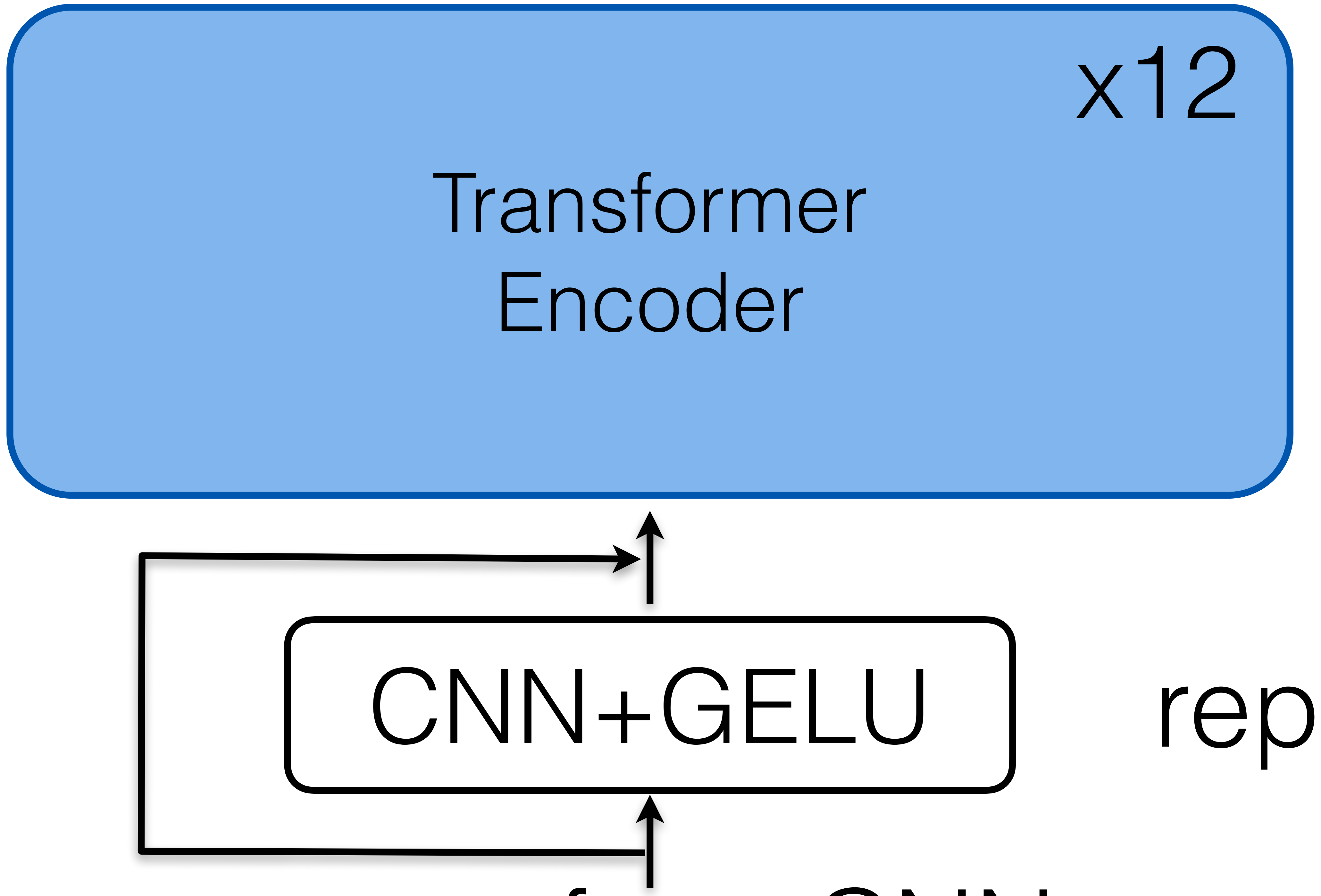
codebookG

one frame vector from CNN

How to obtain codebook — Product Quantization

Splitting a vector into equally sized chunks — subvectors,
Assigning each of these subvectors to its nearest *centroid*

Contextual Encoder



x12

Transformer Encoder

CNN+GELU

replacing pos emb.

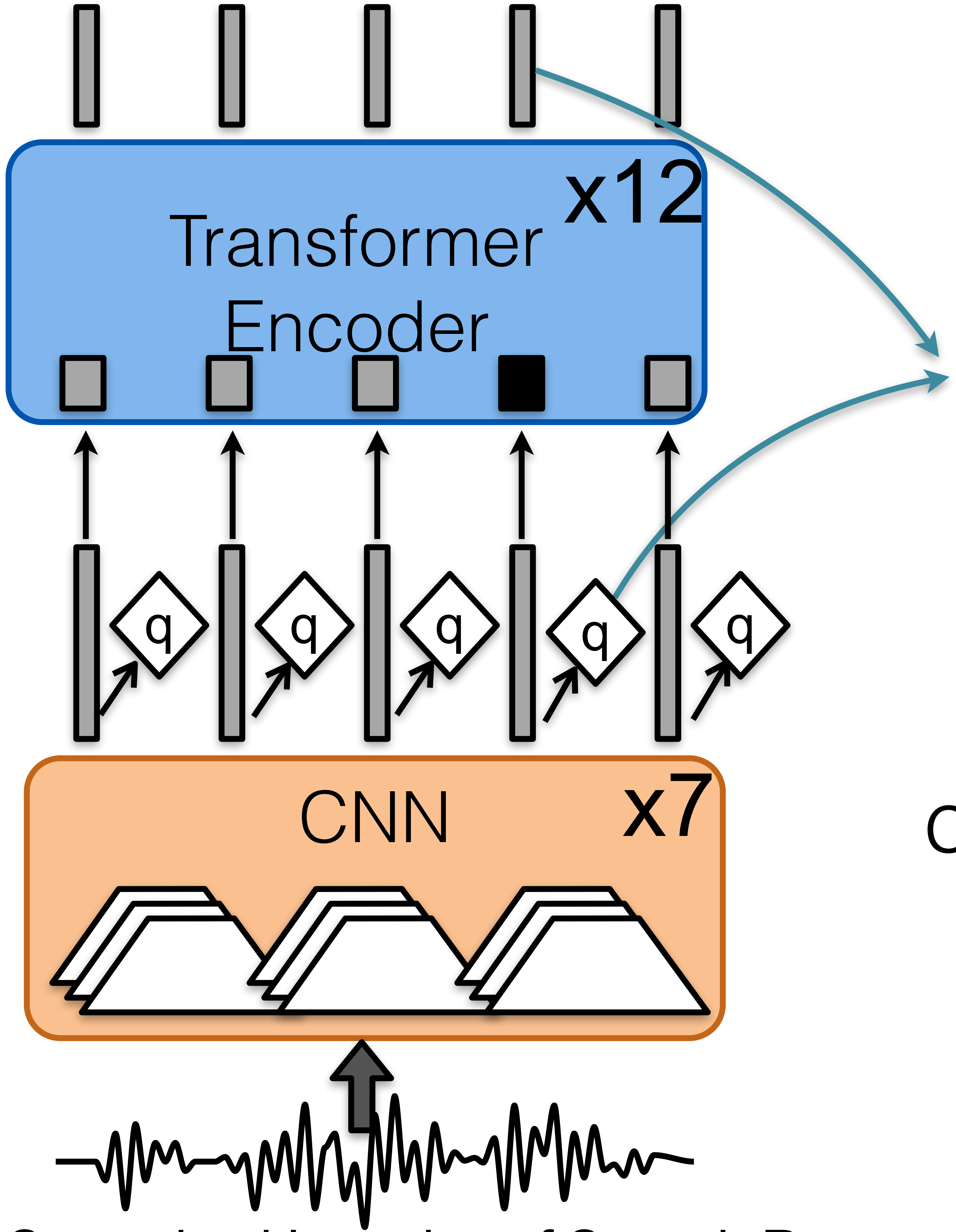
vector from CNN

(before quantization)

Wav2Vec2.0: Contrastive on quantized acoustic state

Masked context during training

Quantized low-level acoustic state, each frame ~ 25ms, stride 20ms



Training data: (audio only)
LibriSpeech 960 hrs

LibriVox 53k hrs

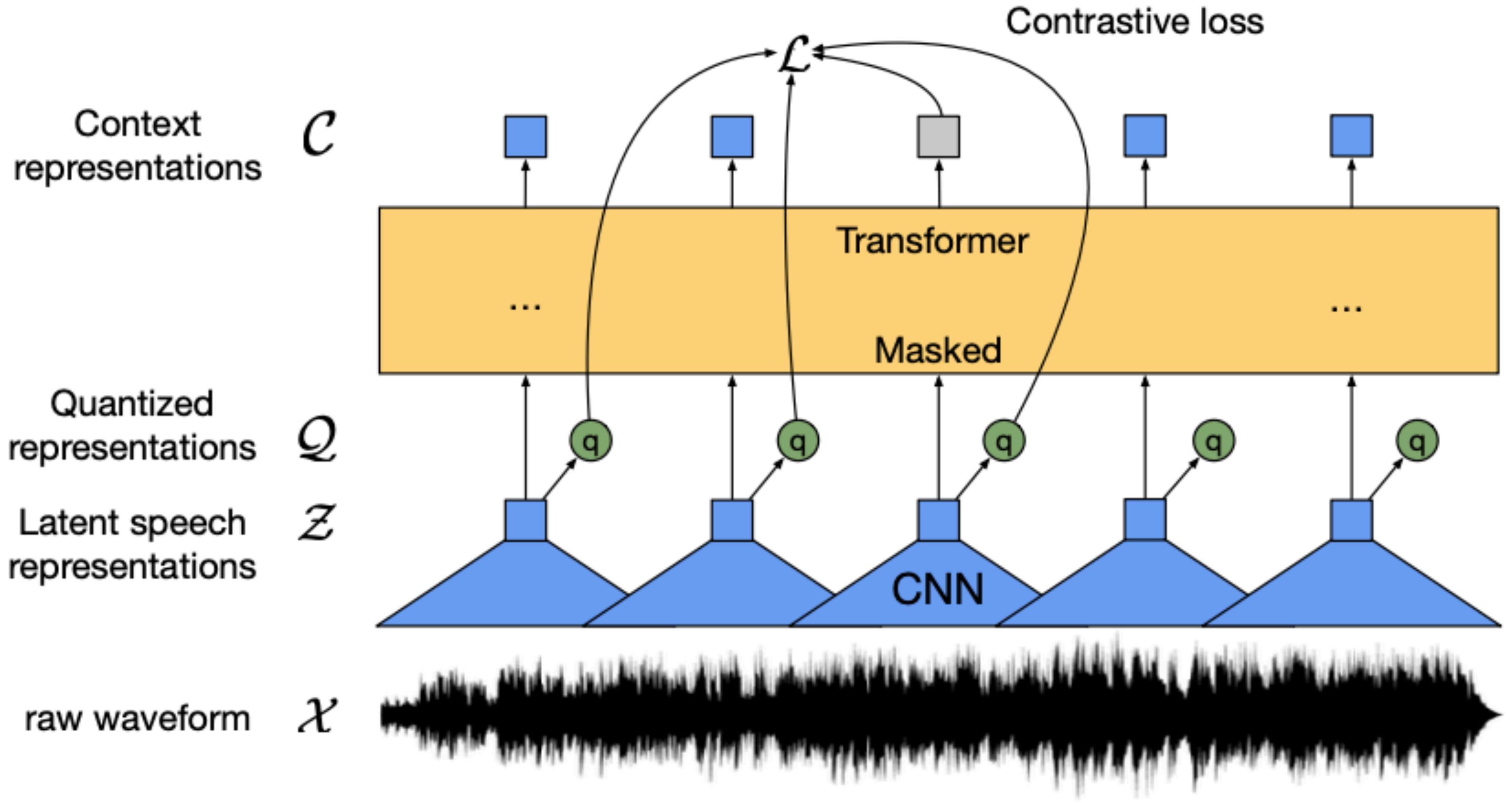
Minimize contrastive loss

$$L = -\sum \log \frac{\exp Sim(c_t, q_t)}{\sum \exp Sim(c_t, q_-)} + \text{penalty}$$

Bring closer masked context and quantized acoustic state

Wav2vec2.0: a Framework for Self-Supervised Learning of Speech Representations [Baevski et al, NeurIPS 2020]

Training Loss


Contrastive loss
$\mathcal{L}$
Context representations $\mathcal{C}$
Transformer
...
...
Masked
Quantized representations $\mathcal{Q}$
q
Latent speech representations $\mathcal{Z}$
CNN
raw waveform $\mathcal{X}$

Cosine similarity

Context representation

Discrete latent speech representation

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

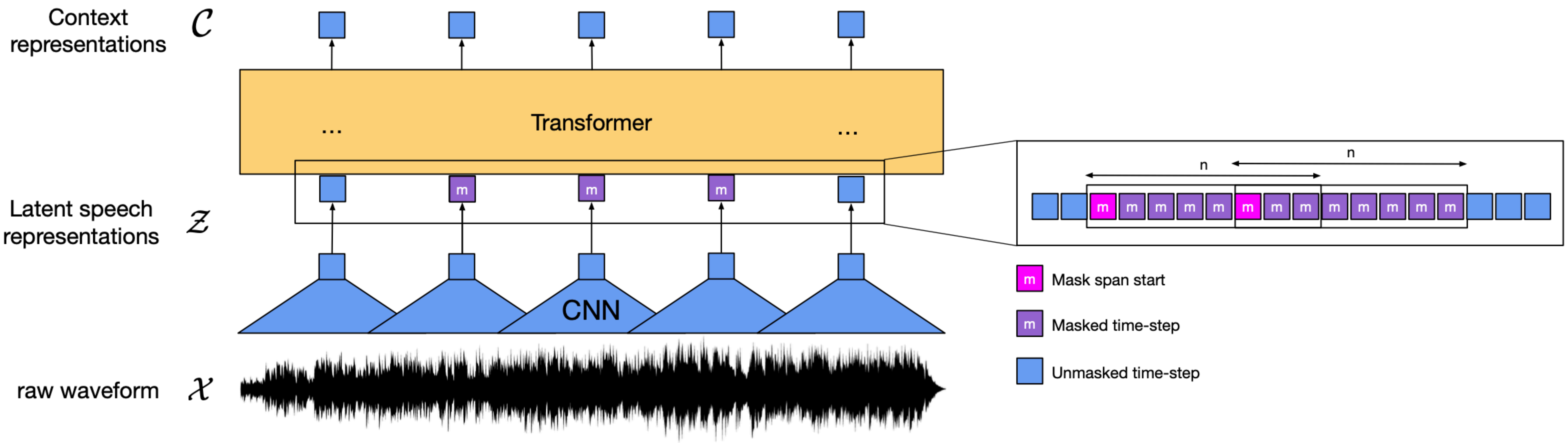
Negative samples

Temperature

Codebook diversity penalty to encourage more codes to be used

Masking

- Sample starting points for masks without replacement, then expand to 10 frames
    - span can overlap
    - for a 15s sample, ~49% of frames masked with an avg span of 300ms



Model Setup

- Wav2vec2 base:
  - 12 Transformer layers, d=768, d_ffn=3072, #heads=8
  - 16 groups
  - rel pos emb cnn kernel size 128

- Wav2vec2 large:
  - 24 Transformer layers, d=1024, d_ffn=4096, #heads=16

Training

- LibriSpeech: 960 hours of English speech (just audio)

- LibriVox (LV-60k): about 53k hours of audio for book reading

- Wav2Vec2 base:
    - each sample is cropped with length 250k (=15.6s)
    - total batch size: 1.6 hours on 64 V100 GPUs

- Wav2Vec2 Large:
    - each sample is cropped with length 320k (=20s)
    - total batch size: 2.7hours on 128 V100 GPUs.

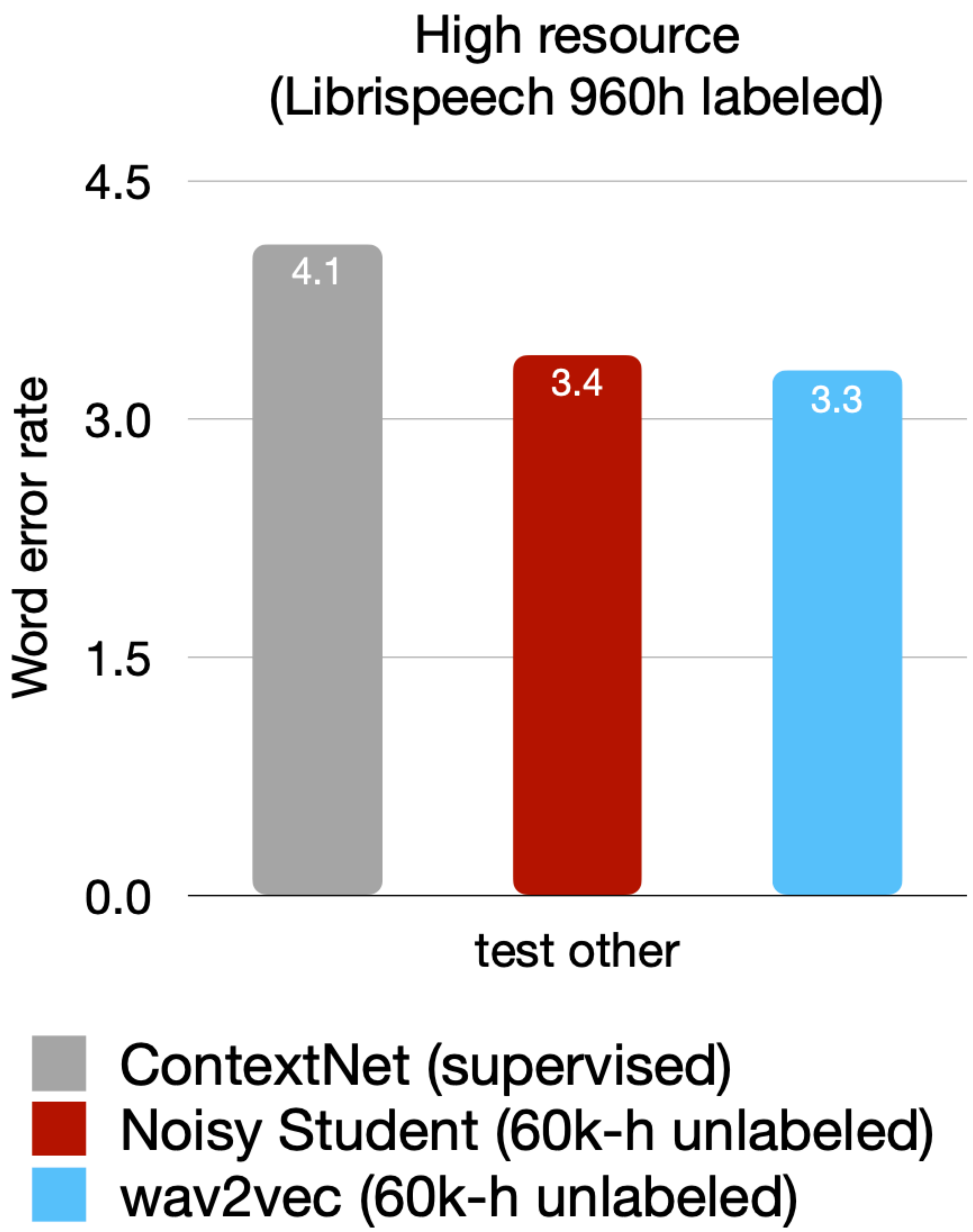
Fine-tuning

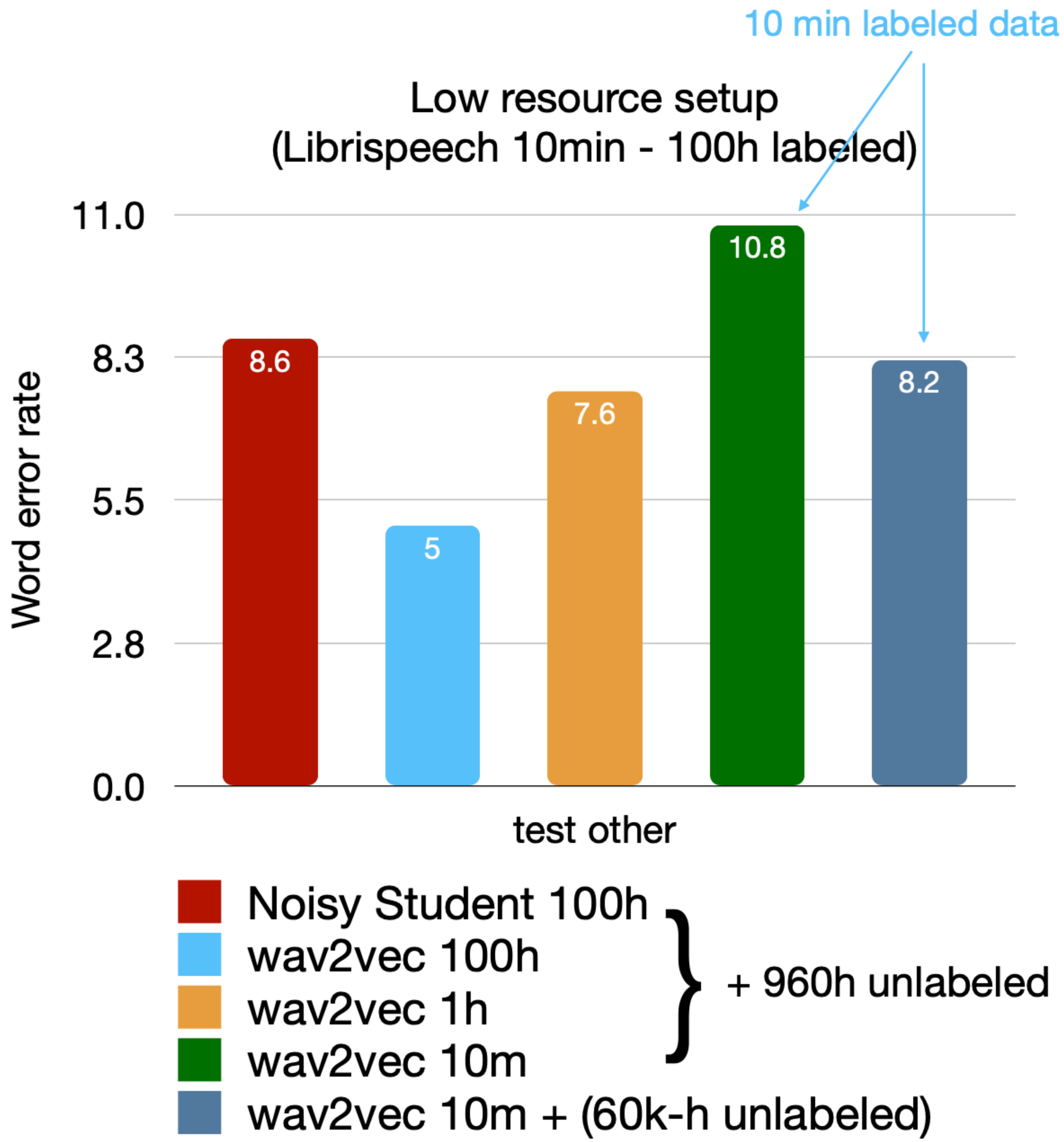
- Add a single linear projection on top into target vocab and train with CTC loss with a low learning rate (CNN encoder is not trained).

- Use modified SpecAugment in latent space to prevent early overfitting

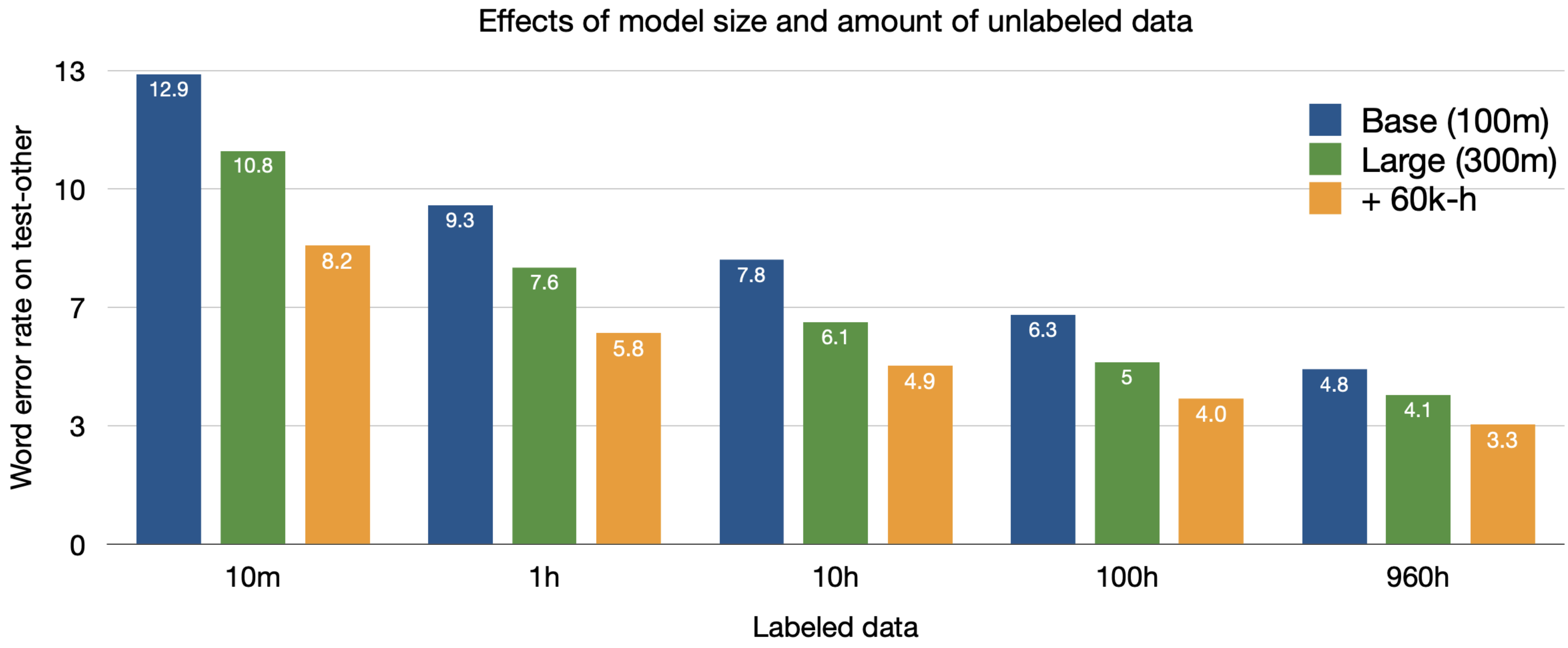
- Uses wav to letter generation with the official 4gram LM and Transformer LM

Wav2Vec2 Results

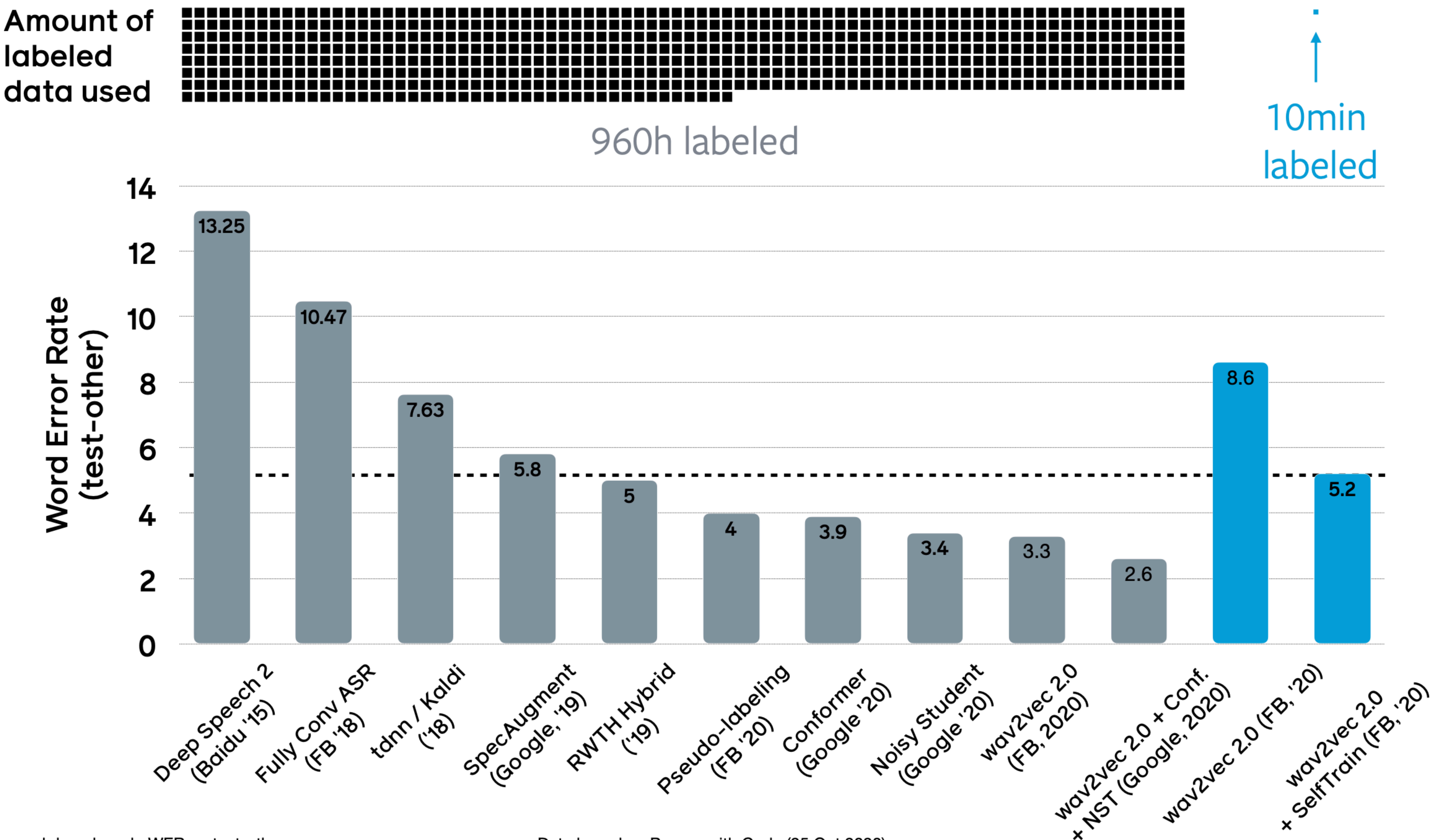




Effects of Model size and raw data



Overall ASR results



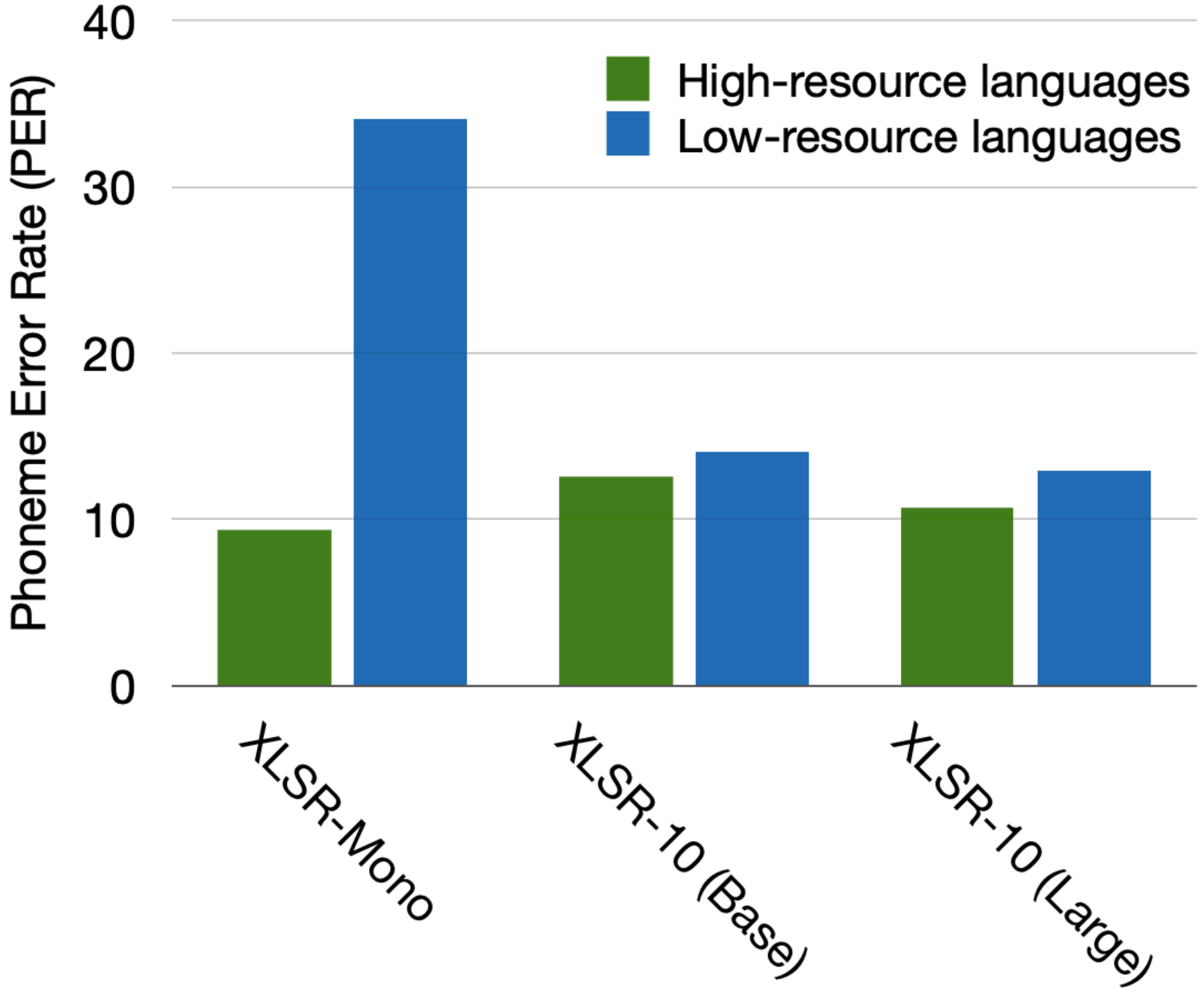
_ibrispeech benchmark, WER on test-other

Data based on Papers with Code (25 Oct 2020)

XLSR: Multilingual Wav2Vec2

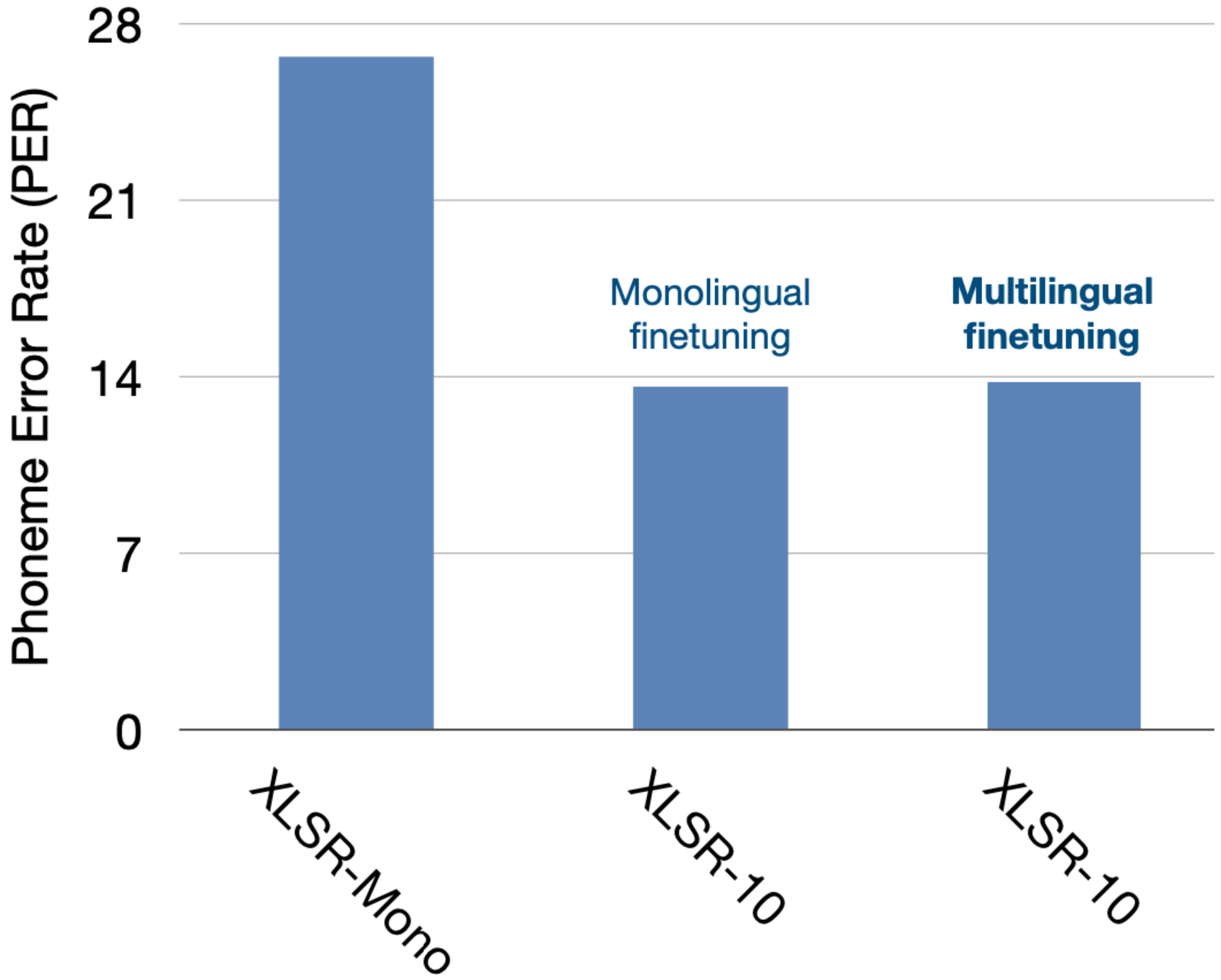
Cross-lingual transfer

CommonVoice results:



Multilingual fine-tuning

CommonVoice results:



Summary

- Self-supervised pre-training with audio data only

- Wav2Vec2 Model: CNN+Transformer

- construct the frames with reasonable size (25ms) and sliding (20ms)
  - proper design of CNNs

- Masked training with contrastive loss on quantized representation

Language in 10

TTS Code in Notebook

- https://github.com/lileicc/FastSpeech2

- https://www.cs.cmu.edu/~leili/course/11737mnlp23fa/code/tts/run_tactron2.ipynb