# wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli

Facebook AI

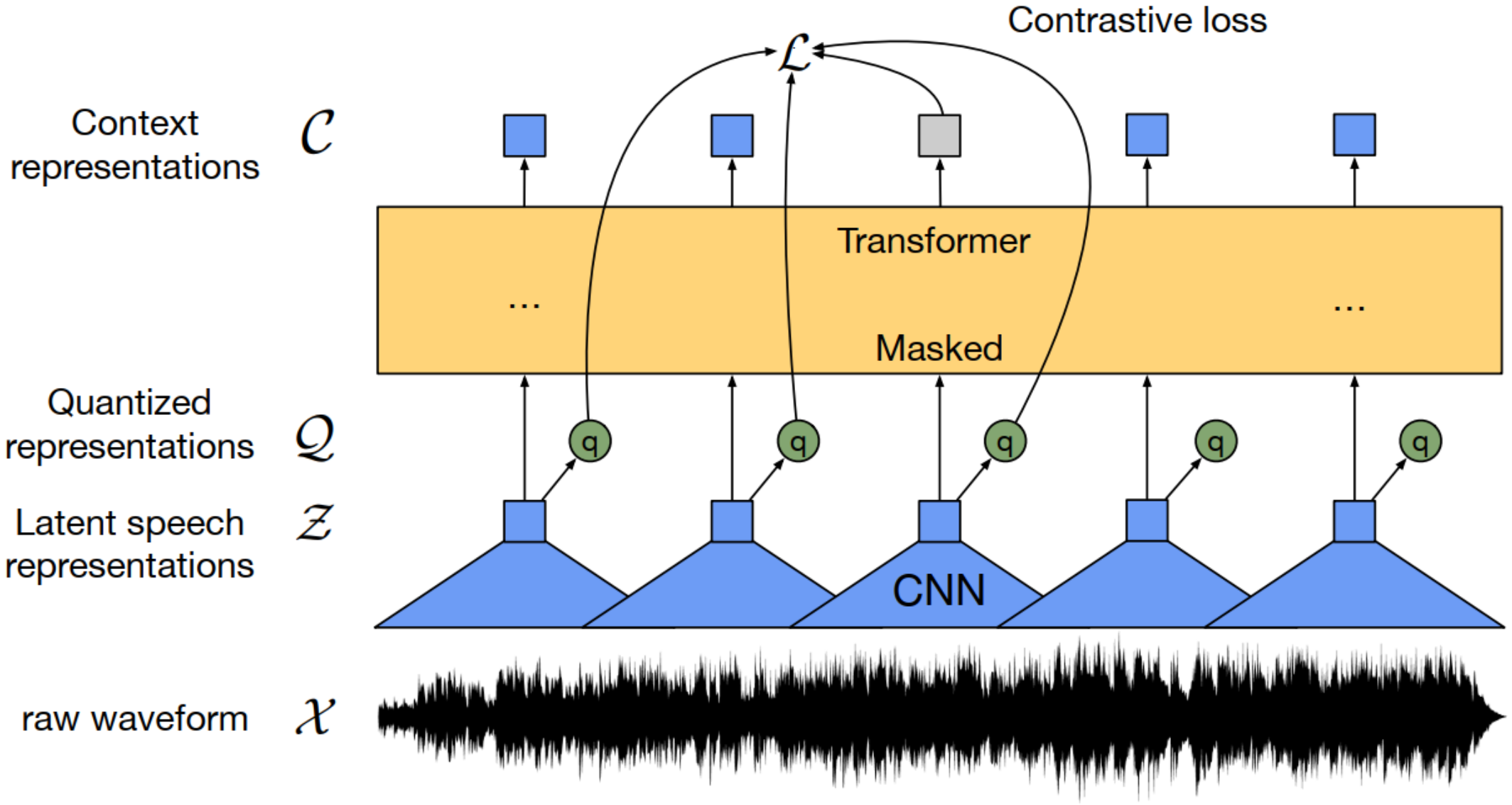
NeurIPS 2020

- Current speech recognition systems require thousands of hours of transcribed speech to reach acceptable performance

- Collecting large amount of labeled data is not available for most of 7,000 languages spoken worldwide

- Self-supervised learning has been very successful for natural language processing and computer vision

- Propose wav2vec 2.0, an end-to-end method for learning discrete speech units as well as contextualized representations from unlabeled speech


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

1. **Feature encoder**

   - Blocks of temporal convolution + layer normalization + GELU activation function

2. **Quantization module**

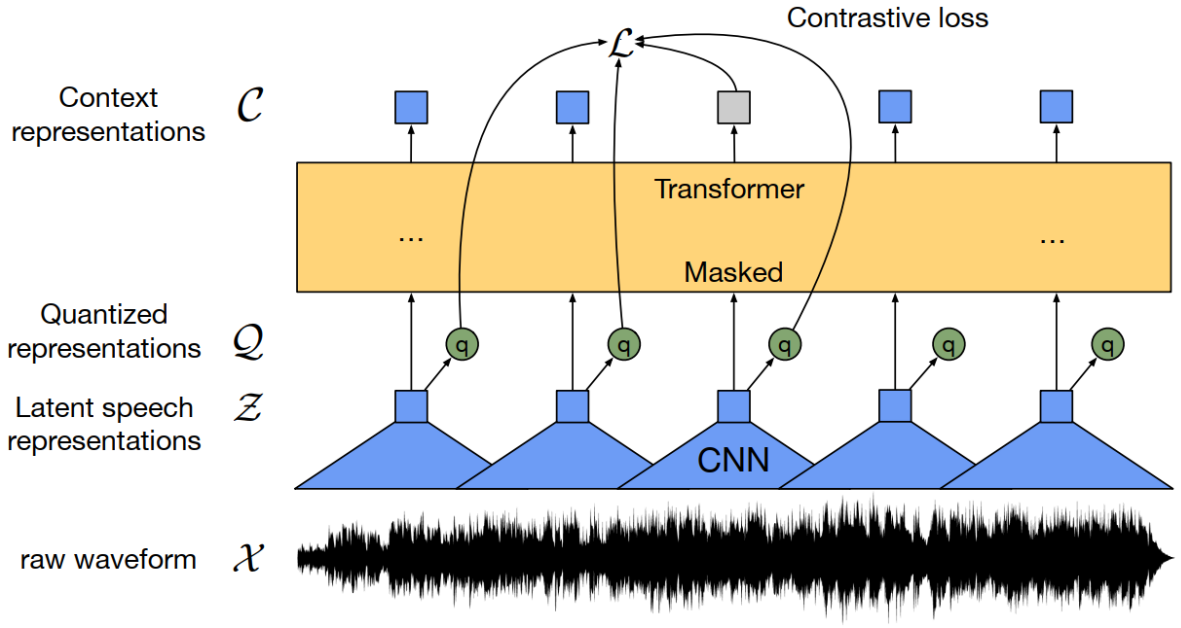
   - Discretize the encodings to a finite set of speech representations via **product quantization**
   - Concatenate $G$ chosen entries from each of $G$ groups containing $V$ entries
   - Gumbel softmax: probabilities for choosing the $v$-th entry in group $g$

$$p_{g,v} = \frac{\exp(l_{g,v} + n_v)/\tau}{\sum_{k=1}^{V} \exp\left(l_{g,k} + n_k\right)/\tau}$$

   where $\tau$ is temperature factor, $n = -\log(-\log(u))$, $u$ are uniform samples from $\mathcal{U}(0,1)$

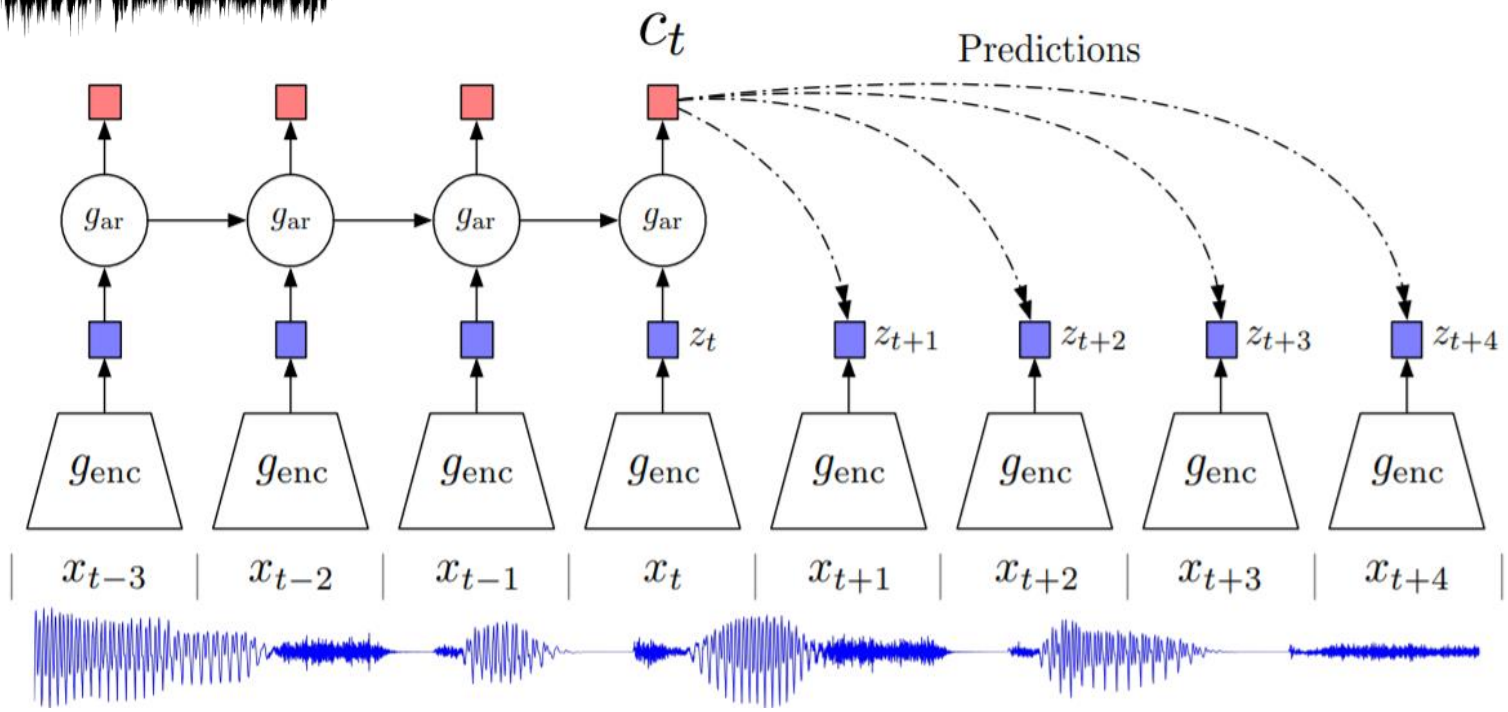
3. **Context network**

   - Convolutional layer + GELU + Transformer encoders




← wav2vec 2.0

CPC →



## 1. Masking

- Randomly mask a proportion of $p$ time steps in encodings and their subsequent $M$ time steps

## 2. Objective

- Contrastive loss: given context network output $\boldsymbol{c}_t$ over masked time step $t$, identify the true quantized $\boldsymbol{q}_t$ from itself and $K$ distractors uniformly sampled from other masked time steps

$$\mathcal{L}_m = -\log \frac{\exp(sim(\boldsymbol{c}_t, \boldsymbol{q}_t)/k)}{\sum_{\tilde{q}\sim Q_t} \exp(sim(\boldsymbol{c}_t, \tilde{\boldsymbol{q}})/k)},$$

  where cosine similarity $sim(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^T\boldsymbol{b}/\|\boldsymbol{a}\|\|\boldsymbol{b}\|$, $k = 0.1$ is a temperature factor

- Diversity loss: maximize the entropy of averaged softmax distribution to **encourage the equal use** of the $V$ entries in each of the $G$ groups

$$\mathcal{L}_d = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}$$

- **Fine-tuning for ASR and phoneme representation**
  - Add a randomly initialized linear projection layer on top of the context network
  - Vocabulary: 29 characters + 1 word boundary token for LibriSpeech dataset, 39 classes for TIMIT
  - SpecAugment
  - Minimize CTC loss

- **Language model (LM) and decoding**
  - 4-gram LM
  - Transformer LM: 20 blocks, model dimension 1,280, FFN dimension 6,144, 16 attention heads

- **Datasets**
  - Pre-training: LibriSpeech-960 hours (LS-960), LibriVox-53.2k hours (LV-60k)
  - Fine-tuning: LS-960, LibriSpeech-100 hours (LS-100), Libri-light (train-10h, train-1h, train-10min), TIMIT-5 hours
- **Model configurations**
  - Feature encoder: 7 blocks with temporal convolution of 512 channels, strides (5, 2, 2, 2, 2, 2, 2), kernel widths (10, 3, 3, 3, 3, 2, 2)
  - Quantization module: $G = 2$, $V = 320$
  - BASE: 12 Transformer blocks, model dimension 768, FFN dimension 3,072, 8 attention heads
  - LARGE: 24 Transformer blocks, model dimension 1,024, FFN dimension 4,096, 16 attention heads

- **Word Error Rates (WER)**
  - [42]: State-of-the-art method
  - LARGE model surpasses [42] with 1-hour labeled data
  - BASE vs. LARGE on LS-960
  - LARGE LS-960 vs. LV-60k

| Model | Unlabeled data | LM | dev | | test | |
|---|---|---|---|---|---|---|
| | | | clean | other | clean | other |
| **10 min labeled** | | | | | | |
| Discrete BERT [4] | LS-960 | 4-gram | 15.7 | 24.1 | 16.3 | 25.2 |
| BASE | LS-960 | 4-gram | 8.9 | 15.7 | 9.1 | 15.6 |
| | | Transf. | 6.6 | 13.2 | 6.9 | 12.9 |
| LARGE | LS-960 | Transf. | 6.6 | 10.6 | 6.8 | 10.8 |
| | LV-60k | Transf. | 4.6 | 7.9 | 4.8 | 8.2 |
| **1h labeled** | | | | | | |
| Discrete BERT [4] | LS-960 | 4-gram | 8.5 | 16.4 | 9.0 | 17.6 |
| BASE | LS-960 | 4-gram | 5.0 | 10.8 | 5.5 | 11.3 |
| | | Transf. | 3.8 | 9.0 | 4.0 | 9.3 |
| LARGE | LS-960 | Transf. | 3.8 | 7.1 | 3.9 | 7.6 |
| | LV-60k | Transf. | 2.9 | 5.4 | 2.9 | 5.8 |
| **10h labeled** | | | | | | |
| Discrete BERT [4] | LS-960 | 4-gram | 5.3 | 13.2 | 5.9 | 14.1 |
| Iter. pseudo-labeling [58] | LS-960 | 4-gram+Transf. | 23.51 | 25.48 | 24.37 | 26.02 |
| | LV-60k | 4-gram+Transf. | 17.00 | 19.34 | 18.03 | 19.92 |
| BASE | LS-960 | 4-gram | 3.8 | 9.1 | 4.3 | 9.5 |
| | | Transf. | 2.9 | 7.4 | 3.2 | 7.8 |
| LARGE | LS-960 | Transf. | 2.9 | 5.7 | 3.2 | 6.1 |
| | LV-60k | Transf. | 2.4 | 4.8 | 2.6 | 4.9 |
| **100h labeled** | | | | | | |
| Hybrid DNN/HMM [34] | - | 4-gram | 5.0 | 19.5 | 5.8 | 18.6 |
| TTS data augm. [30] | - | LSTM | | | 4.3 | 13.5 |
| Discrete BERT [4] | LS-960 | 4-gram | 4.0 | 10.9 | 4.5 | 12.1 |
| Iter. pseudo-labeling [58] | LS-860 | 4-gram+Transf. | 4.98 | 7.97 | 5.59 | 8.95 |
| | LV-60k | 4-gram+Transf. | 3.19 | 6.14 | 3.72 | 7.11 |
| Noisy student [42] | LS-860 | LSTM | 3.9 | 8.8 | 4.2 | 8.6 |
| BASE | LS-960 | 4-gram | 2.7 | 7.9 | 3.4 | 8.0 |
| | | Transf. | 2.2 | 6.3 | 2.6 | 6.3 |
| LARGE | LS-960 | Transf. | 2.1 | 4.8 | 2.3 | 5.0 |
| | LV-60k | Transf. | 1.9 | 4.0 | 2.0 | 4.0 |

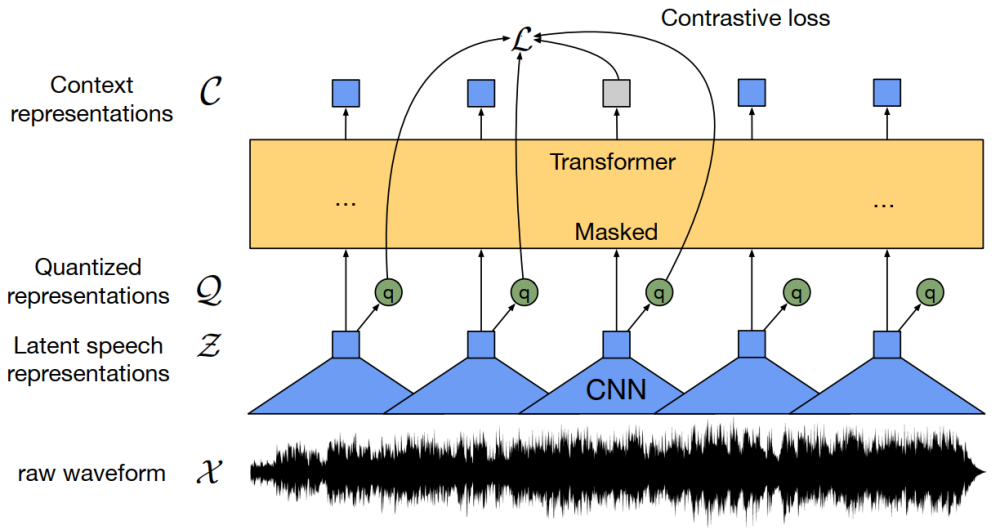
- **WER on LS-960 labeled data**
  - Weaker base model: LARGE -from scratch vs. ContextNet [17]
  - Vocabulary mismatch AM-char; LM-word
  - No data balancing as used in [42]
  - The proposed method and [42] are complimentary

| Model | Unlabeled data | LM | dev | | test | |
|---|---|---|---|---|---|---|
| | | | clean | other | clean | other |
| **Supervised** | | | | | | |
| CTC Transf [51] | - | CLM+Transf. | 2.20 | 4.94 | 2.47 | 5.45 |
| S2S Transf. [51] | - | CLM+Transf. | 2.10 | 4.79 | 2.33 | 5.17 |
| Transf. Transducer [60] | - | Transf. | - | - | 2.0 | 4.6 |
| ContextNet [17] | - | LSTM | 1.9 | 3.9 | 1.9 | 4.1 |
| Conformer [15] | - | LSTM | 2.1 | 4.3 | 1.9 | 3.9 |
| **Semi-supervised** | | | | | | |
| CTC Transf. + PL [51] | LV-60k | CLM+Transf. | 2.10 | 4.79 | 2.33 | 4.54 |
| S2S Transf. + PL [51] | LV-60k | CLM+Transf. | 2.00 | 3.65 | 2.09 | 4.11 |
| Iter. pseudo-labeling [58] | LV-60k | 4-gram+Transf. | 1.85 | 3.26 | 2.10 | 4.01 |
| Noisy student [42] | LV-60k | LSTM | 1.6 | 3.4 | 1.7 | 3.4 |
| **This work** | | | | | | |
| LARGE - from scratch | - | Transf. | 1.7 | 4.3 | 2.1 | 4.6 |
| BASE | LS-960 | Transf. | 1.8 | 4.7 | 2.1 | 4.8 |
| LARGE | LS-960 | Transf. | 1.7 | 3.9 | 2.0 | 4.1 |
| | LV-60k | Transf. | 1.6 | 3.0 | 1.8 | 3.3 |

- **Phoneme Error Rates (PER) on TIMIT**
  - Obtains state-of-the-art with a reduction of 23%/29% over the next best result on the dev/test sets

| | dev PER | test PER |
|---|---|---|
| CNN + TD-filterbanks [59] | 15.6 | 18.0 |
| PASE+ [47] | - | 17.2 |
| Li-GRU + fMLLR [46] | – | 14.9 |
| wav2vec [49] | 12.9 | 14.7 |
| vq-wav2vec [5] | 9.6 | 11.6 |
| **This work (no LM)** | | |
| LARGE (LS-960) | 7.4 | 8.3 |

- **Quantization of the context network input and the targets**
  - BASE pre-trained on LS-960
  - Continuous inputs enable better context representations
  - Quantized targets lead to more robust training




| | avg. WER | std. |
|---|---|---|
| Continuous inputs, quantized targets (Baseline) | 7.97 | 0.02 |
| Quantized inputs, quantized targets | 12.18 | 0.41 |
| Quantized inputs, continuous targets | 11.18 | 0.16 |
| Continuous inputs, continuous targets | 8.58 | 0.08 |

## Contributions

- Presents wav2vec 2.0, a self-supervised learning framework for speech pre-training
- WER of 4.8/8.2 on LibriSpeech with 10 minutes of labeled data
- Outperforms the previous best result with 100 times less labeled data on the clean 100 hour LibriSpeech setup
- Achieves a new state of the art on full LibriSpeech-960 benchmark for noisy speech (test-other)

# Thank you for watching !

Presenter: Wenxin Hou