Improved Language Identification Through Cross-Lingual Self-Supervised Learning

Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh,
Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, Michael Auli

Meta AI

Outline

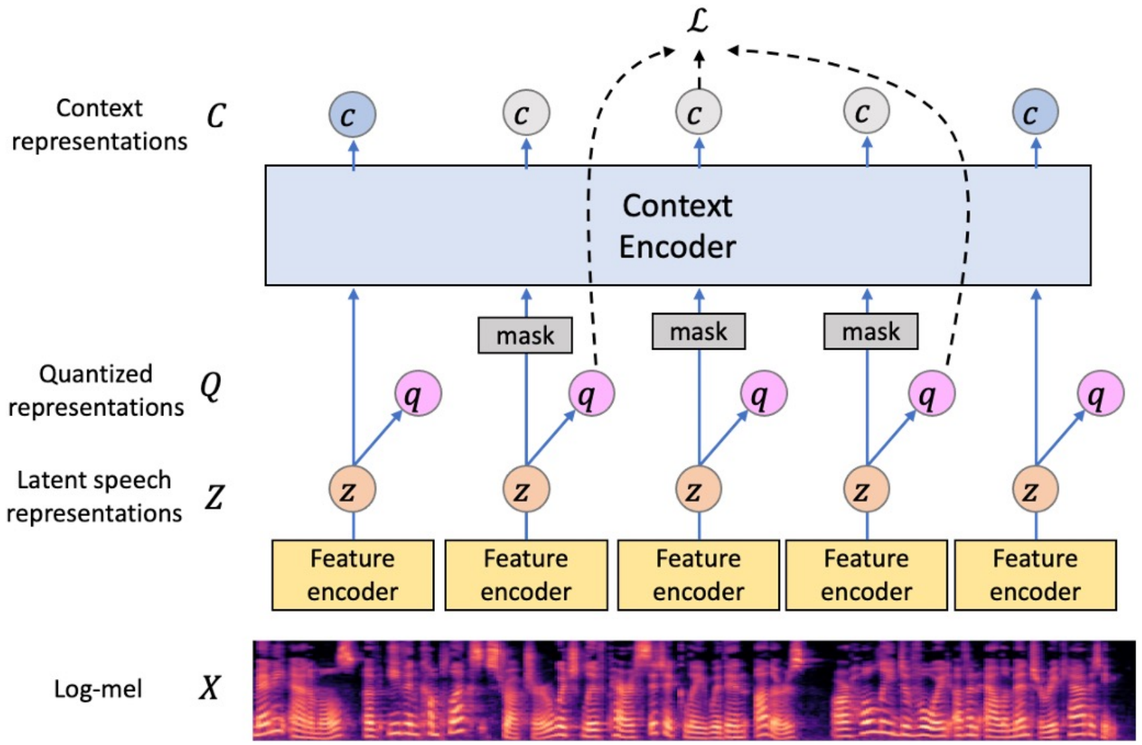
1. Background
2. Log-Mel Wav2Vec
3. Cross-lingual Speech Representation (XLSR)
4. LID Finetuning
5. Experimental setup
6. Results
7. Conclusion

Background

- Language identification (LID) predicts which language is being spoken given a speech utterance and routes the speech to the correct ASR service.
- Problem & prior works:
  - Modern LID are trained with large amounts of labeled data.
  - Self-supervised learning (e.g., Wav2Vec 2.0) can leverage unlabeled data.
  - Prior study [1] shows Wav2Vec 2.0 trained on English data could improve LID.
- Proposed:
  - Extend the study using cross-lingual self-supervised (XLSR) to improve LID accuracy.
  - Explore different 1) content aggregation strategies, 2) pruning layer size.

[1] Exploring wav2vec 2.0 on Speaker Verification and Language Identification, Fan et al. Interspeech 2021

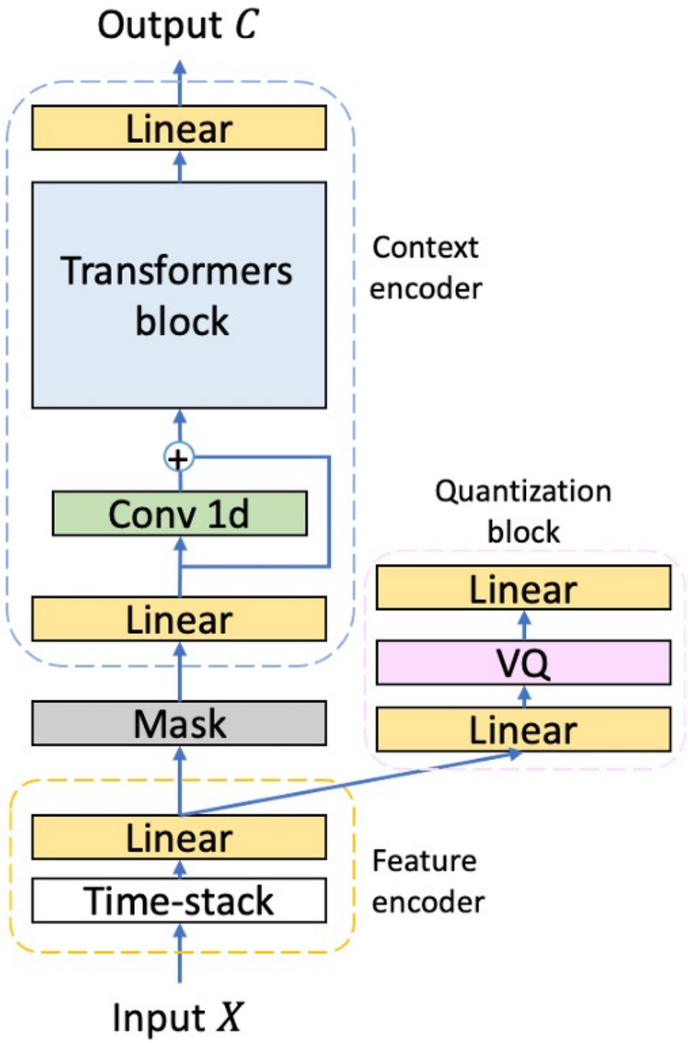
Wav2Vec 2.0



- Wav2Vec 2.0 is a self-supervised learning trained with audio only.
- It consists of 3 main components
  - Feature encoder: extract latent speech representation Z
  - Quantizer: learn contextualized representation Q from continuous Z
  - Context encoder: learn high-level speech representation
- Loss function:

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

Log-Mel Wav2Vec

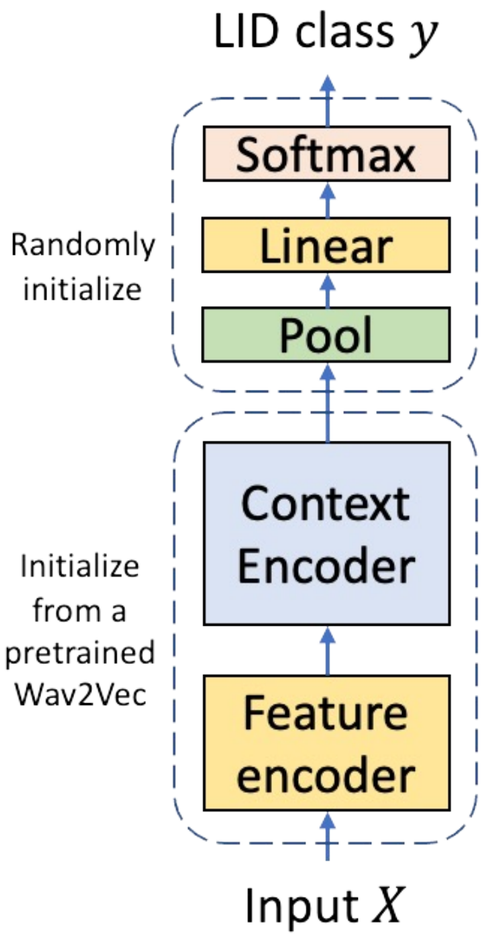


- Instead of using raw-waveform, we replace the input feature into log-mel spectrogram.
- We also modify the Wav2Vec architecture by replacing stack of 1D convolution layers with time-stacking + a linear layer.
- By using this modification, we could reduce the memory usage and improve running time to support large-scale training more efficiently.

Cross-lingual Speech Representation (XLSR)

- XLSR is a multilingual Wav2Vec 2.0, trained altogether with various unlabeled speech from different languages.
- In this paper, since we don’t have any language metadata on the unlabeled speech, we train our XLSR without any data rebalancing.

LID Finetuning



- We initialized the bottom part of the LID classifier with a pre-trained Wav2Vec.
- We explored several pooling operations such as:
  - Mean pooling $o = \sum_{t=1}^{T} c_t / T$
  - Std. dev pooling $o = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (c_t - \mu)^2}$
  - Self-attention pooling

$$o = \sum_{t=1}^{T} softmax\left(w_2 GELU\left(W_1 c^{\mathrm{T}}\right)\right) c_t$$

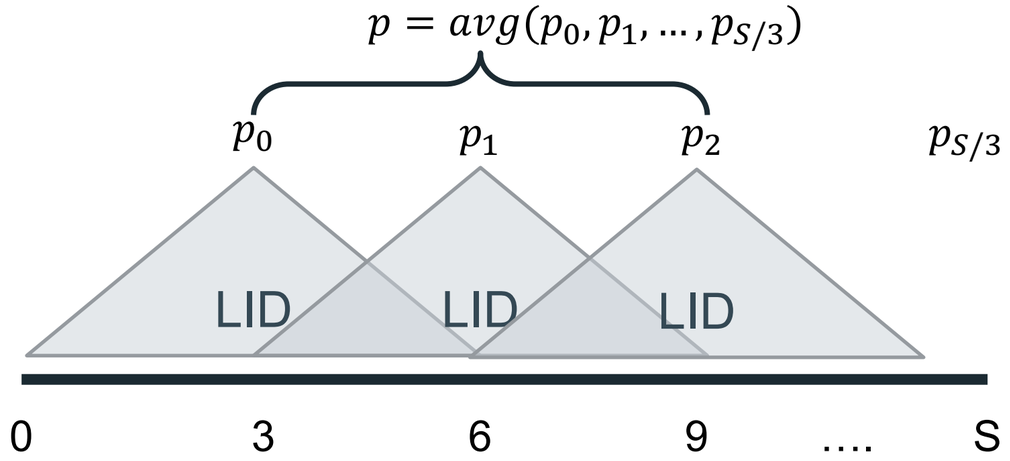
  - Concat [CLS] token in the 1st index and set $o = c_1$
- Loss criterion: cross-entropy

Experimental Setup

- Pre-training XLSR
  - We are using 6.3 million hours unlabeled speech with different languages and conditions (e.g., clean, noisy, etc).
  - Log-mel Wav2Vec configuration:
    - Feature encoder: 4x time-stack stride + linear layer
    - Context encoder: 24 Transformers layer (d_in = 1024, d_ffn=4096)
    - Quantization module: Gumbel VQ (320 codebooks, 2 groups)
- Finetuning
  - We finetune with LID dataset consisted of 26 languages

Experimental Setup (cont.)

- Evaluation
  - Inference with 6 seconds audio & 3 seconds step size and we average the language probability across multiple segments to predict the class.

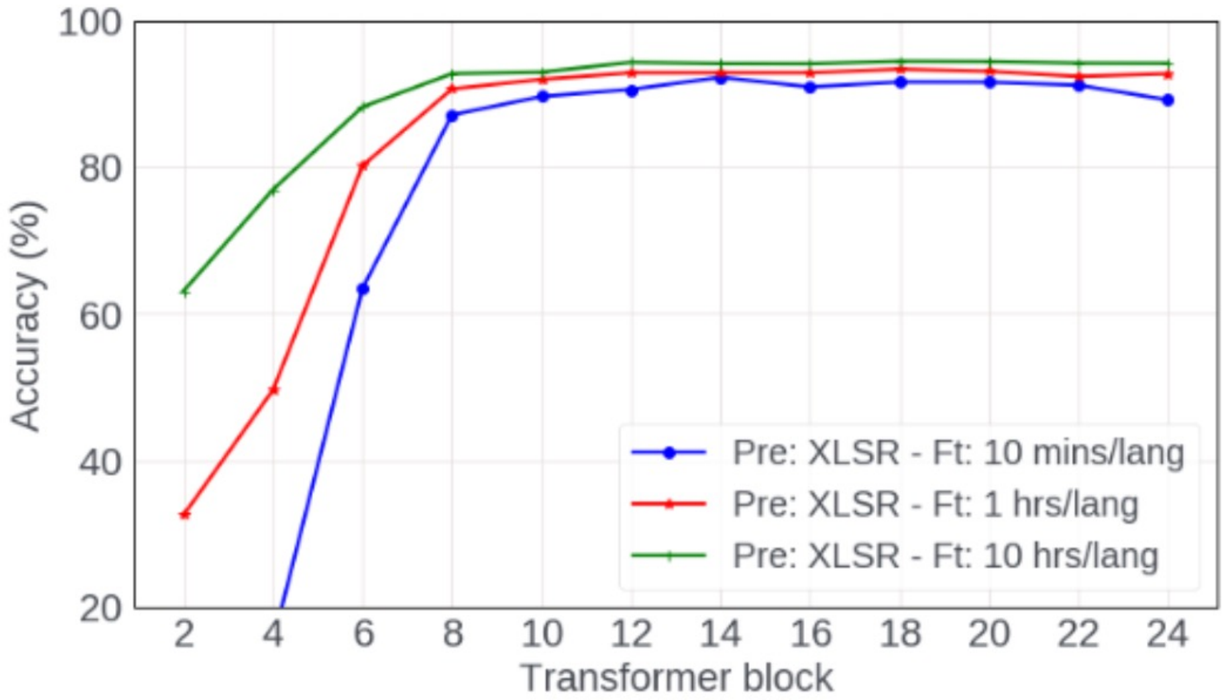


Exp 1: LID accuracy in different setups

Lbl. / lang	Pre-training	Test Accuracy (%)			
		0-6s	6-18s	18-∞s	Overall
10 min	None	7.1	9.5	10.6	9.6
	w2v2 En	71.3	73.1	76.1	74.2
	XLSR	85.4	88.8	90.8	**89.2**
1 hour	None	20.2	25.2	29.5	26.5
	w2v2 En	79.3	85.9	89.3	86.5
	XLSR	87.2	92.5	94.8	**92.8**
10 hours	None	48.3	61.9	71.8	64.5
	w2v2 En	86.8	93.3	95.6	93.4
	XLSR	88.2	94.3	96.1	**94.2**
100 hours	None	72.2	84.9	90.7	86.7
	w2v2 En	89.5	95.7	97.3	95.5
	XLSR	90.3	95.9	97.2	**95.7**

- We compare 3 different models: 1) without pre-training, 2) pre-trained with Wav2Vec English data, 3) XLSR
- We also compare the accuracy between different amount of labeled data, test data lengths.
- XLSR shows best performance overall.

Exp 2: Pruning context Transformers layers



- We tried to prune the layers from context encoders.
- Our results shows that we could prune up to 2/3 layers (reduce params from 300m -> 100m) and still maintains the same accuracy.

Exp 3: Pooling strategy

Aggregation strategy	Accuracy (%)			
	0-6s	6-18s	18-∞s	Overall
Max	86.6	92.7	94.8	92.8
Mean+Max+Min	88.1	92.9	94.7	93.0
Mean+Max	88.5	93.1	94.8	93.2
Mean+Std	84.2	90.9	93.4	91.1
[CLS] Token	85.4	91.4	93.9	91.7
Self Attention	87.0	92.0	94.1	92.2

- We tried different pooling on 1hr/lang setup. Mean+Max pooling shows the best accuracy despite its simplicity.

Conclusion

- We showed that cross-lingual pre-trained model are particularly effective for low-resource setups, especially when little labeled data is available.
- Using only 10 minutes per languages, an XLSR-based LID could achieve 89.2% on 26 languages setup.
- From our pruning study, we could prune 2/3 of context encoder layers and still maintain its accuracy.
- Simple pooling strategy such as max+mean works well in this system.