Unsupervised Pre-Training for Vietnamese Automatic Speech Recognition in the HYKIST Project

Bachelor Thesis Colloquium

Le Duc Khai
dle@i6.informatik.rwth-aachen.de
RWTH Aachen University

08.12.2022

Examiner:	Prof. Dr. rer. nat. Ilya E. Digel, FH Aachen
Co-Examiner:	Christoph M. Lüscher, M.Sc., RWTH Aachen
Advisor:	Sen. Prof. Dr.-Ing. Hermann Ney, RWTH Aachen
Co-Advisor:	PD Dr. rer. nat. Ralf Schlüter, RWTH Aachen

HYKIST project

- Doctor speaks German, patient speaks Arabic or Vietnamese. Interpreter speaks German and Arabic, or German and Vietnamese
- Goal: is to use **Automatic Speech Recognition** (ASR) and **Machine Translation** (MT) to assist interpreters $\rightarrow$ minimize time to look up medical terms, or interpreters' mistakes.

Current problems

- Lack of training data: 6h (HYKIST) and 219h (in-house data)
- Available Vietnamese datasets and pretrained models (XLSR-53) are sampled at 16 kHz, while ours at 8kHz
- Domain mismatch: General telephone conversations (in-house data) vs. medical-focused conversations (HYKIST)
- Accented speech: Local accents in in-house data vs foreign-born accents in HYKIST
- Conversational speech: natural flow of speech, stuttering, emotional speakers...
- Linguistic aspect of Vietnamese medical terms: Vietnamese language is monosyllabic (having only one syllable per word), while medical terms are loan words (polysyllabic but with Vietnamese pronunciation)

Research questions

- Roughly 100M speakers but no work has deeply investigated unsupervised pretraining for Vietnamese yet.
- No work has applied unsupervised pretraining to difficult low-resource medical tasks.
- No work has investigated the use of unsupervised pretraining methods for telephone speech directly on the 8kHz signal without resampling.
- The analysis of regularization for a medical ASR system has never been presented.

Related work

- **RWTH ASR Systems for LibriSpeech: Hybrid vs Attention** [Lüscher et al. 19]
  - Hybrid ASR setup achieving SOTA on Librispeech dataset
- **wav2vec 2.0: A framework for self-supervised learning of speech representations** [Baevski et al. 20]
  - The self-supervised learning framework demonstrating a good ASR system with limited amounts of labeled data
- **Unsupervised cross-lingual representation learning for speech recognition** [Conneau et al. 20]
  - Release a large multilingual model *XLSR-53*
  - Multilingual pretraining outperforms monolingual pretraining
- **Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training** [Hsu et al. 21]
  - Pretraining on multiple domains to analyze the effect on target language
- **Wav2vec-aug: Improved self-supervised training with limited data** [Sriram et al. 22]
  - Use data augmentation for pretrained data to boost ASR accuracy
- **Deja-vu: Double feature presentation and iterated loss in deep transformer networks** [Tjandra et al. 20]
  - Propose the intermediate loss to boost accuracy

Hybrid ASR system [Lüscher et al. 19]

- **Bayes theorem**: The relation w^* between the acoustic observations x_1^T whose length is T and the most likely word sequence to be recognized w_1^N:

$$w^* = \arg\max_{w_1^N} \underbrace{p(x_1^T | w_1^N)}_{\text{acoustic model}} \cdot \underbrace{p(w_1^N)}_{\text{language model}} \tag{1}$$

- **Gammatone feature extraction** [Schlüter et al. 07]: Hanning window $\rightarrow$ (10th root or log) compression $\rightarrow$ DCT-based cepstral decorrelation $\rightarrow$ normalization
- **Hidden Markov Model (HMM)**:

$$p(x_1^T | w_1^N) = \sum_{[s_1^T]} \prod_{t=1}^{T} p(x_t, s_t | s_{t-1}, w_1^N) = \sum_{[s_1^T]} \prod_{t=1}^{T} \underbrace{p(s_t | s_{t-1}, w_1^N)}_{\text{transition prob.}} \cdot \underbrace{p(x_t | s_t, s_{t-1}, w_1^N)}_{\text{emission prob.}} \tag{2}$$

Hybrid ASR system [Lüscher et al. 19]

- **Context-Dependent Phone**: generate triphone or allophone labels
- **CART** [Beulen and Ney 98]: To cluster triphones $\rightarrow$ reduce the number of labels
- **Baum–Welch algorithm**: To find the best alignment between acoustic observations and transcriptions:
  1. Maximization
  2. Expectation
  3. Get back to step 1

Hybrid ASR system [Lüscher et al. 19]

- **GMM/HMM system**: HMM to model the transition between phones and the corresponding observable, GMM to model the emission probabilities
- **DNN/HMM system**: DNN to model the posterior probability $p(a_{s_t}|x_t)$ discriminatively. The emission probability in the HMM can be calculated such that:

$$p(x_t|a_{s_t}) = \frac{p(a_{s_t}|x_t)p(x_t)}{p(a_{s_t})}. \tag{3}$$

  The probability $p(a_{s_t})$ can be estimated as the relative frequency of a_{s_t}. The probability $p(x_t)$ can be removed.

Hybrid ASR system [Lüscher et al. 19]

- **Language modeling (LM)**: 4-gram count based LM, using Kneser-Ney Smoothing algorithm [Kneser and Ney 95]. Full-words used in the first-pass decoding [Beck et al. 19].
- **Decoding**: With the help of the Viterbi approximation:

$$w_1^N = \arg\max_{N, w_1^N} p\left(\prod_{n=1}^{N} p(w_n | w_{n-m}^{n-1}) \cdot \max_{[s_1^T]} \prod_{t=1}^{T} p(x_t, s_t | s_{t-1}, w_1^N)\right), \tag{4}$$

  Beam search (AM and LM pruning) used to only focus on the most promising predicted words at each time step [Ortmanns et al. 97].
- **Recognition performance**: The Word-Error-Rate (WER) indicator is used:

$$\text{WER} = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{Reference words}} \tag{5}$$

wav2vec 2.0 [Baevski et al. 20]

- **Self-supervised learning**:

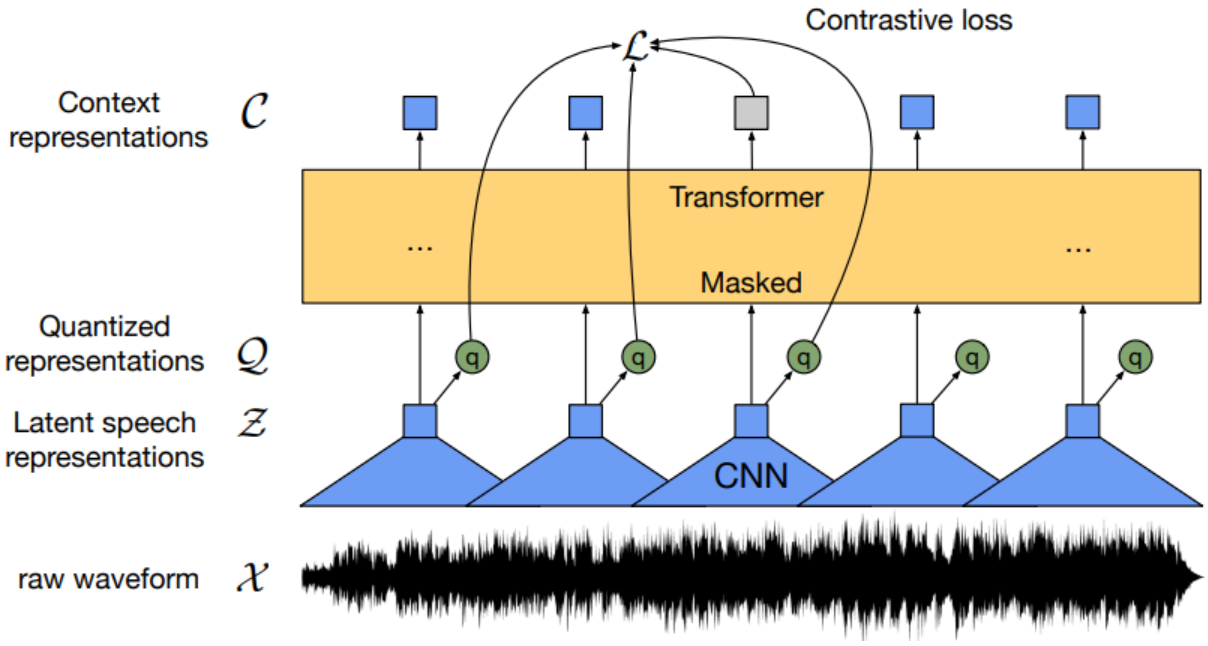
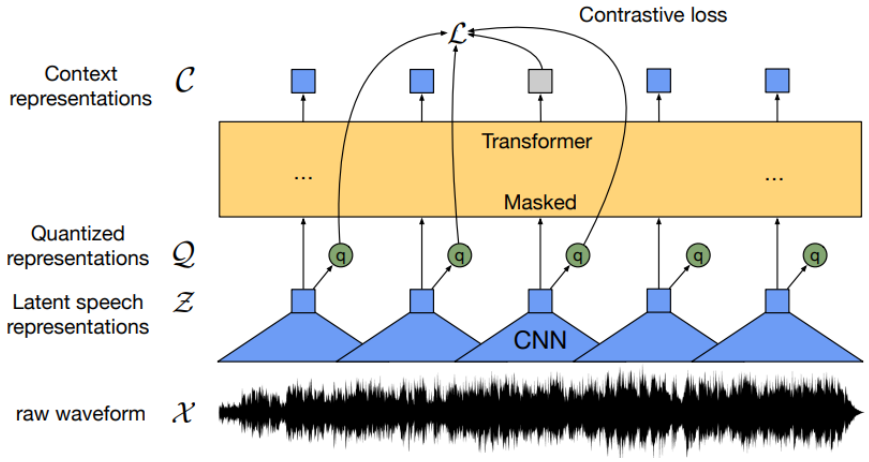


Illustration of wav2vec 2.0 framework which jointly learns contextualized speech representations and an inventory of discretized speech units.

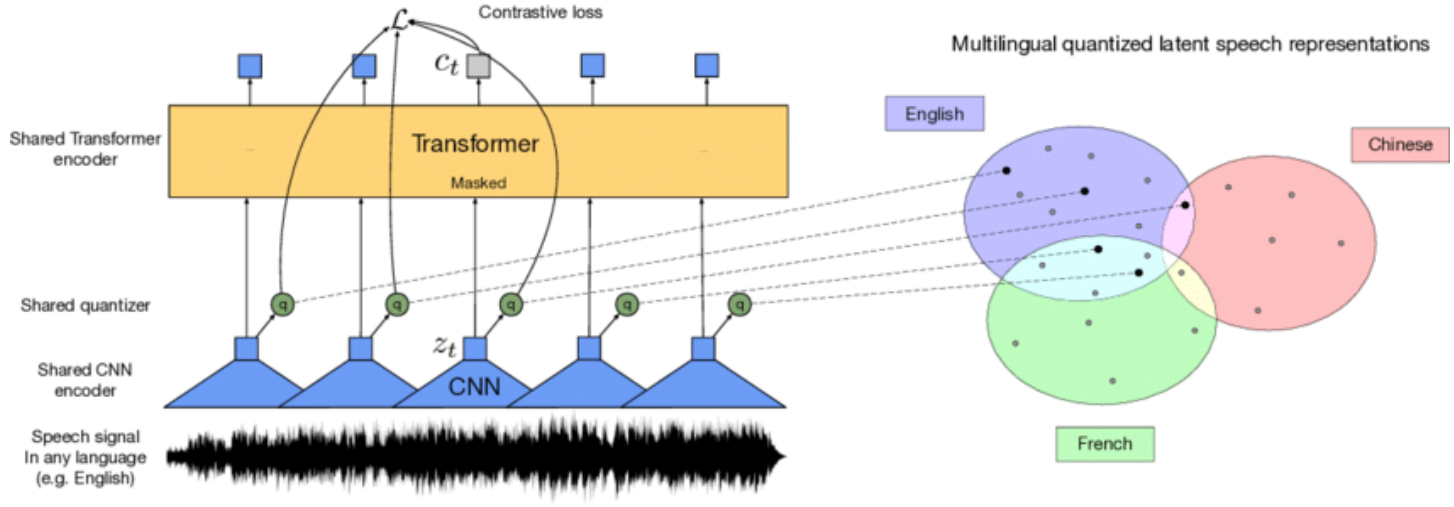
wav2vec 2.0 [Baevski et al. 20]

- **Feature encoder**: Blocks of temporal convolution, layer normalization [Ba et al. 16], and the GELU activation function [Hendrycks and Gimpel 18].
- **Contextualized representations with Transformers**: A context network that uses the Transformer architecture [Vaswani et al. 17].
- **Contrastive learning**: Get a context representation c_t for a masked latent representation z_t in order to guess the proper quantized representation q_t among alternative quantized representations.



Unsupervised cross-lingual representation learning (XLSR) [Conneau et al. 20]

- **XLSR**: By extending wav2vec 2.0 to the multilingual setting, cross-lingual learning seeks to create models that use data from other languages to improve performance.
- **In-domain Match Level**: Determined by the similarity between recording conditions, naturalness and conversational topics.
- **Diversity Level**: of the multilingual dataset is higher than the monolingual one because the first represents more learnable phonemes helpful to target language.



Training data

- Acoustic training data:

Language	Dataset	Usage	# Spks	Hours	Domain	In-domain match	Diversity level
Arabic	In-house	pretr.	3379	786	Tel., Conv.	Medium	Medium
German	In-house	pretr.	1723	177	Tel., Conv.	Medium	Medium
Vietnamese	In-house	pretr., finetu.	2240	219	Tel., Conv.	Medium	Medium
	HYKIST	adapt	1	1	Tel., Conv., Med.	High	Low
		dev	3	3			
		test	2	2			
	YouTube	pretr.	-	1.204	Read books	Low	
Multi	In-house*	pretr.	7342	1.182	Tel., Conv.	Medium	High
	XLSR-53	pretr.	-	56.000	Various	Low	

*The multilingual in-house training dataset is the combination of the Arabic, German and Vietnamese ones listed above. Domain: Telephone (Tel.), Conversational (Conv.), Medical (Med.). Sampling rate: 8kHz

Training data

- Data for the large out-of-domain recognition: Read speech datasets resampled from 16kHz to 8kHz:
  - CommonVoice Vietnamese (17h of noisy recordings) [Ardila et al. 20]
  - VIVOS (15h of clean recordings) [Luong and Vu 16]
- Lexicon and language model:

# words	vocab	dev		test	
in train	size	OOV	PPL	OOV	PPL
500M	11k	0.1%	67	0.2%	69

4-gram LM for Vietnamese

Acoustic model

- Supervised-only models: GMM, BLSTM (25M parameters), Transformer (90M)
- Models using unsupervised pretraining (317M by full architecture *Large*, 115M parameters by cut-off $Large_{1\text{-}8}$ and 95M by *Base* architecture):
  - XLSR-53
  - Models pretrained from scratch with custom data
  - Continued pretraining with XLSR-53 as initialization
- Regularization:
  - Unsupervised pretraining: speed perturbation [Ko et al. 15], random pitch perturbation [Chen et al. 20], reverberation perturbation [Peddinti et al. 15]
  - Finetuning: SpecAugment [Park et al. 19], Intermediate Cross-entropy Loss (ICE Loss), Intermediate Focal Loss (IF Loss), L2 Regularization [Krogh and Hertz 91], On-off Regularization

Experimental results

Supervised baselines

AM	WER [%]	
	Hykist dev	Hykist test
GMM	62.2	59.7
BLSTM	32.9	38.4
Transformer	31.0	35.1

WER [%] for supervised-only baselines on Vietnamese HYKIST data [Lüscher et al. 22].
Models are trained on the monolingual in-house data.

Experimental results

Monolingual pretraining

Pre-training			Fine-tuning	WER [%]	
Data	Hours	Epochs	Epochs	Hykist dev	Hykist test
None	-	None	33	32.1	36.6
Viet. in-house (sup. pret.)	219	33		31.9	36.9
Viet. in-house (unsup. pret.)		100	26	31.4	33.4
Aug. Viet. in-house	1168	300		31.0	32.3
Viet. YT				29.8	35.2
Viet. in-house + YT				25.3	27.2

All fine-tunings use the $Large_{1\text{-}8}$ architecture and are trained until full convergence on Vietnamese in-house data and the recognition is done on HYKIST. All pre-trainings have been done with **random initialization**. Pre-training data "None" in the 3rd row means fine-tuning from scratch with wav2vec 2.0 $Large_{1\text{-}8}$ architecture. In the experiment "supervised pretraining" in 4th row, we mimic the learning rate curve of the experiment "unsupervised pretraining" in 5th row. The rest pretraining schedules are unsupervised. The chosen number of pretraining epochs is the best checkpoint.

Experimental results

Monolingual pretraining

- WERs of from scratch (32.1% and 36.6%) and in-house supervised pretraining (31.9% and 36.9%) reduce to (31.4% and 33.4%) of in-house pretraining
  → On **wav2vec 2.0** architecture, the unsupervised pretraining helps
- WERs reduce from 31.4% and 33.4% to 31.0% and 32.3%
  → Data augmentation for pretraining is helpful
- WERs of Youtube data (29.8% and 35.2%) vs. in-house data (31.4% and 33.4%)
  → Having more data is not always helpful, if the domain mismatch is larger and the data has less speakers
- WERs by 25.3% and 27.2% when combining the in-house and YT data (the best result using solely monolingual data)
  → Diversity of domains and speakers is necessary

Experimental results

Multilingual pretraining

Pre-training				WER [%]	
Init	Data	Hours	Epochs	Hykist dev	Hykist test
random	Viet. in-house + YT	1168	300	25.3	27.2
	Multilingual in-house			26.8	28.7
XLSR-53$_{1\text{-}8}$	None	-	None	27.6	31.9

All fine-tunings use the *Large*$_{1\text{-}8}$ architecture and are trained until full convergence on Vietnamese in-house data and the recognition is done on HYKIST. The 2nd model is pretrained on our multilingual in-house dataset, and the 3rd uses *XLSR-53*$_{1\text{-}8}$ to directly finetune on Vietnamese in-house data. The chosen number of pretraining epochs is the best checkpoint.

Multilingual pretraining

- WERs of multilingual (26.8% and 28.7%) vs. monolingual combination (25.3% and 27.2%)
  $\rightarrow$ Monolingual data with multiple domains outperforms multilingual data with high domain-match and speaker diversity.
- WERs by 27.6% and 31.9% when directly finetuning with *XLSR-53*
  $\rightarrow$ May be due to the absence of 8kHz data in the pretraining of *XLSR-53*

XLSR-53 as pretraining initialization

Pre-training				WER [%]	
Arch.	Data	Hours	Epochs	Hykist dev	Hykist test
$Large_{1\text{-}8}$	None	-	None	27.6	31.9
	Viet. in-house	219	25	27.6	29.5
Large			100	26.2	29.0
$Large_{1\text{-}8}$	Viet. YT	1168	100	24.3	28.1
	Viet. in-house + YT			24.5	27.2
	Multilingual in-house		50	23.9	27.4

Table for unsupervised pre-training with the public *XLSR-53* model as initialization. All finetunings use the $Large_{1\text{-}8}$ architecture. The 1st model is the direct finetuning on Vietnamese in-house data, and the remaining models use *XLSR-53* as initialization for pretrainings (full model *Large* or cut-off model $Large_{1\text{-}8}$). The chosen number of pretraining epochs is the best checkpoint.

XLSR-53 as pretraining initialization

- WERs of in-house Vietnamese data (31.4% and 33.4% in table Monolingual pretraining) vs. continued pretraining on in-house Vietnamese (27.6% and 29.5%) vs. direct finetuning with $\textit{XLSR-53}_{\text{1-8}}$ (27.6% and 31.9%).
  $\rightarrow$ Continued pretraining using *XLSR-53* model outperforms the pretraining using random initialization and the direct finetuning using *XLSR-53*.
- WERs of full *XLSR-53* reduce from 27.6% and 29.5% to 26.2% and 29.0%
  $\rightarrow$ If the resource usage is neglected, the *Large* model should be chosen for better accuracy instead of the cut-off $\textit{Large}_{\text{1-8}}$.
- WERs of multilingual pretraining reduce from 26.8% and 28.7% (table Multilingual pretraining) to 23.9% and 27.4%, those for YT reduce from 29.8% and 35.2% (table Monolingual pretraining) to 24.3% and 28.1%, those for in-house+YT reduce from 25.3% and 27.2% (table Monolingual pretraining) to 24.5% and 27.2%
  $\rightarrow$ Continued pretraining helps both the monolingual and the multilingual scenario. However, the continued pretraining on less diverse data benefits more from the diverse and multilingual data.

Experimental results

Comparison to supervised baselines

AM	Init	Pre-training		WER [%]	
		Data	Hours	Hykist dev	Hykist test
Transformer	random	None	-	31.0	35.1
wav2vec 2.0		Viet. in-house	219	31.4	33.4
		Viet. YT	1168	29.8	35.2
		Viet. in-house + YT		25.3	27.2
	XLSR-53$_{1\text{-}8}$			24.5	27.2
		Multilingual in-house		23.9	27.4

All fine-tunings use the *Large*$_{1\text{-}8}$ architecture and are trained until full convergence on Vietnamese in-house data and the recognition is done on HYKIST. The 1st model is the supervised-only training using Transformer. The 2nd, 3rd and 4th models are pretrained on specific data using random initializaton. The 5th and 6th models are continued pretraining methods (using *XLSR-53*$_{1\text{-}8}$ as initialization).

Comparison to supervised baselines

- WERs of in-house pretraining (31.4% and 33.4%) vs. YT data (29.8% and 35.2%) are higher than Transformer baseline (31.0% and 35.1%).
  $\rightarrow$ wav2vec 2.0 unsupervised pretraining does not always outperform the Transformer supervised-only approach, especially when the pretrained data is not diverse enough.
- The best results are (24.5 % and 27.2%) on monolingual data and (23.9% and 27.4%) on multilingual data.
  $\rightarrow$ Continued pretraining should be used regardless of sampling rate mismatch to gain the most benefits in terms of accuracy.

Encoder comparison

Architecture	Pre-training		WER [%]	
	Data	Hours	Hykist dev	Hykist test
Base	None	-	35.8	39.9
$Large_{1\text{-}8}$			35.0	40.7
Base	Viet. in-house	219	30.2	33.3
$Large_{1\text{-}8}$			31.5	33.4
Base	Multilingual in-house	1168	26.2	28.8
$Large_{1\text{-}8}$			26.8	28.7

Comparison for *Base* and $Large_{1\text{-}8}$ using different pretraining schedules: no pretraining, pretraining on in-house data and on multilingual data. All fine-tunings are done until full convergence on Vietnamese in-house data and the recognition is done on HYKIST.

- WERs between *Base* and $Large_{1\text{-}8}$ fluctuate
  $\rightarrow$ We recommend the use of *Base* (97M parameters) in order to keep the performance competitive to $Large_{1\text{-}8}$ (118M) while reducing the number of trainable parameters.

Experimental results

Initialization comparison

Architecture	Init. scheme	Pretraining			WER [%]	
		Data	Hours	Epochs	Hykist dev	Hykist test
Base	Kaiming Init.	Viet. in-house	0.01	1	30.6	35.2
$Large_{1\text{-}8}$					31.8	35.7
Base	Glorot Init.	None	-	None	35.8	39.9
$Large_{1\text{-}8}$					35.0	40.7

WERs for architecture *Base* and $Large_{1\text{-}8}$ using 2 different initialization schemes:
Kaiming Initialization (in Fairseq framework [Ott et al. 19]) and Glorot Initialization (in RETURNN framework [Doetsch et al. 17]).

- For super short pretraining (1 epoch pretraining on only 0.01h of data), the results outperform those of raw waveform from scratch for both *Base* and $Large_{1\text{-}8}$ architecture
  $\rightarrow$ In our experiments Kaiming Initialization shows better performance but the small differences in pretraining might lead to a less clear comparison.

Effectiveness of Intermediate Cross-Entropy Loss

- Improvement on HYKIST data: Total improvements for 3/9 experiments of small out-of-domain recognition

Pre-training		With ICE	WER [%]	
Data	Hours		Hykist dev	Hykist test
None	-	No	35.6	40.7
		Yes	33.8	38.1
Viet. YT	1168	No	29.8	35.2
		Yes	27.3	31.5
Viet. in-house + YT		No	25.3	27.2
		Yes	25.1	27.1

Improvements of WERs on HYKIST data between pretraining schedules when applying ICE Loss. All pretrainings use the *Large*$_{1\text{-}8}$ architecture with random initialization and are finetuned until full convergence on Vietnamese in-house data. Only 1 intermediate layer is applied in the middle Transformer block, e.g. position 4 for *Large*$_{1\text{-}8}$ and 6 for *Base* architecture.

Effectiveness of Intermediate Cross-Entropy Loss

- Degradation on HYKIST data: Total degradation for directly finetuning with *XLSR-53*. The rest are partial degradations.

Arch.	Init.	Pre-training		With ICE	WER [%]	
		Data	Hours		Hykist dev	Hykist test
$Large_{1\text{-}8}$	*XLSR-53*	None	-	No	27.9	32.3
				Yes	28.4	33.3
	None	Viet. in-house	219	No	30.4	33.4
				Yes	29.1	33.7
	XLSR-53			No	25.5	29.1
				Yes	25.2	29.2
Base	None			No	30.2	33.3
				Yes	29.7	33.4
$Large_{1\text{-}8}$		Multiling. in-house	1168	No	26.8	28.7
				Yes	25.5	29.4
	XLSR-53	Viet. YT		No	24.3	28.1
				Yes	23.7	28.2

Degradations of WERs on HYKIST data between pretraining schedules when applying ICE Loss. All models are finetuned until full convergence on Vietnamese in-house data.

Effectiveness of Intermediate Cross-Entropy Loss

- Improvement on CommonVoice and VIVOS data: Total improvements for 3/9 experiments of large out-of-domain recognition

Pre-training		With ICE	WER [%]		
Data	Hours		CV dev	CV test	Vivos
None	-	No	20.8	44.7	34.9
		Yes	18.6	42.1	33.1
Viet. YT	1168	No	16.4	34.4	28.7
		Yes	15.6	32.2	27.6
Viet. in-house	219	No	16.4	35.6	31.3
		Yes	16.1	34.8	30.4

Improvements of WERs on CommonVoice and VIVOS between pretraining schedules when applying ICE Loss. All pretrainings use the $Large_{1\text{-}8}$ architecture with random initialization and are finetuned until full convergence on Vietnamese in-house data.

Effectiveness of Intermediate Cross-Entropy Loss

- Degradation on CommonVoice and VIVOS data: Total degradations for continued pretraining on in-house data and for pretraining on in-house + YT data. The rest are partial degradations.

Arch.	Init.	Pre-training		With ICE	WER [%]		
		Data	Hours		CV dev	CV test	Vivos
Large$_{1\text{-}8}$	*XLSR-53*	None	-	No	14.8	32.5	30.3
				Yes	15.8	33.9	30.0
		Viet. in-house	219	No	11.5	29.4	27.2
				Yes	12.3	29.8	27.7
Base	None			No	16.6	35.4	30.9
				Yes	15.4	34.2	31.3
Large$_{1\text{-}8}$		Multiling. in-house	1168	No	15.2	29.7	29.5
				Yes	14.8	30.5	28.8
		Viet. in-house + YT		No	12.9	26.5	21.0
				Yes	13.6	28.2	21.9
	XLSR-53	Viet. YT		No	11.8	28.4	25.6
				Yes	12.3	28.3	25.0

Degradations of WERs on CommonVoice and VIVOS between pretraining schedules when applying ICE Loss.

Effectiveness of Intermediate Focal Loss

- Improvement on HYKIST data:

Arch.	Init.	Pre-training		With IF	WER [%]	
		Data	Hours		Hykist dev	Hykist test
Large$_{1\text{-}8}$	None	None	-	No	35.6	40.7
				Yes	33.0	38.8
		Viet. in-house	219	No	30.4	33.4
				Yes	28.6	33.0
	XLSR-53			No	25.5	29.1
				Yes	24.7	29.1
Base	None			No	30.2	33.3
				Yes	29.0	33.0
Large$_{1\text{-}8}$		Viet. YT	1168	No	29.8	35.2
				Yes	26.1	30.8
		Viet. in-house + YT		No	25.3	27.2
				Yes	24.5	27.1
	XLSR-53	Viet. YT		No	24.3	28.1
				Yes	23.4	28.1

Improvements of WERs on HYKIST data between pretraining schedules when applying IF Loss. All models are finetuned until full convergence on Vietnamese in-house data.

Effectiveness of Intermediate Focal Loss

- Improvement on HYKIST data:
  - 7/9 experiments of IF Loss experience total improvements, compared to only 3/9 experiments of ICE Loss. In addition, all WERs of IF Loss experiments are lower than those of ICE Loss experiments, except the one on HYKIST test set of from scratch training.
    $\rightarrow$ When finetuning and recognizing on the same telephone domain, IF Loss works better than ICE Loss.
  - WERs reduce from (29.8% and 35.2%) to (26.1% and 30.8%) for YT, from (30.4% and 33.4%) to (28.6% and 33.0%) for Vietnamese in-house, from (25.3% and 27.2%) to (24.5% and 27.1%) for in-house+YT.
    $\rightarrow$ Effectiveness of IF Loss decreases when the pretrained data becomes more diverse.

Effectiveness of Intermediate Focal Loss

- Degradation on HYKIST data:

Init.	Pre-training		With IF	WER [%]	
	Data	Hours		Hykist dev	Hykist test
XLSR-53	None	-	No	27.9	32.3
			Yes	28.0	32.8
None	Multiling. in-house	1168	No	26.8	28.7
			Yes	25.2	29.3

Degradations of WERs on HYKIST data between pretraining schedules when applying IF Loss. All pretrainings use the $Large_{1\text{-}8}$ architecture and are finetuned until full convergence on Vietnamese in-house data.

The degradation of *XLSR-53* is rather small. Partial degradation in the multilingual in-house data but the average WER of dev and test set (25.2% and 29.3%) is even lower than the baseline (26.8% and 28.7%).

$\rightarrow$ In a rapid deployment of an ASR system, we recommend the direct use of IF Loss in training without the need of one more training without intermediate loss as a baseline for performance comparison.

Effectiveness of Intermediate Focal Loss

- Improvement on CommonVoice and VIVOS data:

Arch.	Pre-training		With IF	WER [%]		
	Data	Hours		CV dev	CV test	Vivos
$Large_{1\text{-}8}$	Viet. in-house	219	No	16.4	35.6	31.3
			Yes	15.8	34.5	29.6
	None	-	No	20.8	44.7	34.9
			Yes	19.7	43.1	33.9
Base	Viet. in-house	219	No	16.6	35.4	30.9
			Yes	15.9	34.4	30.5
$Large_{1\text{-}8}$	Viet. YT	1168	No	16.4	34.4	28.7
			Yes	14.5	30.9	26.9

Improvements of WERs on CommonVoice and VIVOS between pretraining schedules when applying IF Loss. All models are finetuned until full convergence on Vietnamese in-house data.

Effectiveness of Intermediate Focal Loss

- Improvement on CommonVoice and VIVOS data:
  - The ICE Loss makes 3/9 experiments totally improved (table Improvement on CommonVoice and VIVOS data), while the IF Loss makes 4/9. WERs for IF Loss on 3 read speech datasets are as competitive as ICE Loss. In addition, on the same telephone domain, IF Loss works better than ICE Loss.
    $\rightarrow$ The IF Loss works better than the traditional ICE Loss in all domains.

Effectiveness of Intermediate Focal Loss

- Degradation on CommonVoice and VIVOS data:

Init.	Pre-training		With IF	WER [%]		
	Data	Hours		CV dev	CV test	Vivos
XLSR-53	None	-	No	14.8	32.5	30.3
			Yes	15.4	33.6	30.0
	Viet. in-house	219	No	11.5	29.4	27.2
			Yes	13.0	29.8	27.6
None	Multiling. in-house	1168	No	15.2	29.7	29.5
			Yes	14.5	30.6	28.1
	Viet. in-house + YT		No	12.9	26.5	21.0
			Yes	12.7	28.6	22.1
XLSR-53	Viet. YT		No	11.8	28.4	25.6
			Yes	13.2	29.1	24.5

Degradations of WERs on CommonVoice and VIVOS between pretraining schedules when applying IF Loss. All pretrainings use the $Large_{1\text{-}8}$ architecture and are finetuned until full convergence on Vietnamese in-house data.

- Degradations in experiments pretrained on diverse data
  $\rightarrow$ We recommend IF Loss only for less diverse pretrained data if the domain of finetuning and recognition data are too different.

Studies on Intermediate Focal Loss design

Viet. in-house $Large_{1\text{-}8}$					
Layer	Hykist dev	Hykist test	CV dev	CV test	Vivos
None	30.4	33.4	16.4	35.6	31.3
2	29.4	34.0	**15.5**	35.7	30.0
4	**28.6**	**33.0**	15.8	**34.5**	**29.6**
6	29.1	**33.0**	16.6	34.1	29.9
2,6	29.1	34.1	15.6	35.5	30.7
3,5	29.1	33.3	16.4	35.0	30.1
Viet. in-house *Base*					
None	30.2	33.3	16.6	35.4	30.9
3	**28.7**	33.4	17.0	35.5	**30.1**
6	29.0	33.0	**15.9**	**34.4**	30.5
9	29.5	**32.6**	14.8	34.4	30.5
4,8	29.3	33.5	16.2	35.0	30.3

WERs comparison of IF Loss on different layers between 2 architecture sizes: *Base* and $Large_{1\text{-}8}$. Layer "None" means the baseline (no application of IF Loss).

- Degradation on different single/double layer for $Large_{1\text{-}8}$, mixed results for the *Base*.
  $\rightarrow$ Single IF Loss in the middle network layer yields the best result among variants.

Experimental results

On-off Regularization technique

Off reg.				On reg.		
Epochs	Hykist dev	Hykist test	Continue fine-tuning $\longrightarrow$	Epochs	Hykist dev	Hykist test
0	-	-		33	33.0	38.8
3	45.6	48.6		33	**32.5**	**38.4**
7	43.4	45.9		33	34.3	39.5
10	44.6	47.1		33	34.1	39.5

WERs comparison of On-off Regularization technique over epochs for raw waveform from scratch training. "Off Regularization" stage means training without regularization techniques like Dropout, SpecAugment and IF Loss. After training for some first epochs, the learning rate is reset and the model is preloaded in the "On Regularization" stage (all regularization techniques are turned on). All models are then continued being finetuned until full convergence on Vietnamese in-house data and recognized on HYKIST dataset. The 3rd row (0 epoch for "Off Regularization") is the baseline.

- "Off Regularization" stage for the first 3 epochs and then "On Regularization" stage, WERs reduce from (33.0% and 38.8%) to (32.5% and 38.4%).
- In the future work, we plan to apply the "On-off Regularization" technique to other pretraining schedules.

Combination of L2 regularization and Intermediate Focal Loss

Arch.	Init.	Pre-training		Reg.	WER [%]	
		Data	Hours		Hykist dev	Hykist test
Large$_{1\text{-}8}$	None	Viet. in-house	219	With IF	28.6	33.0
				With IF + L2	28.6	32.9
		None	-	With IF	33.0	38.8
				With IF + L2	31.4	36.4
Base		Viet. in-house	219	With IF	29.0	33.0
				With IF + L2	28.8	32.3
Large$_{1\text{-}8}$	*XLSR-53*			With IF	24.7	29.1
				With IF + L2	24.3	29.2

WERs comparison of the L2 Regularization combination with IF Loss.

- The right parameters for L2 are chosen based on grid-search strategy and different parameters make the WERs vary greatly.
  $\rightarrow$ We recommend L2 regularization should be the last regularization effort in the entire regularization pipeline due to its higher sentitivity to WERs compared to ICE Loss and IF Loss.

Conclusion

Overall results

1. We use various acoustic encoder topologies to present supervised-only baselines for hybrid HMM.
2. Unsupervised pretraining is especially effective when diverse pretraining data is used, e.g. data on multiple domains, multi-speaker data, augmented data... Also, monolingual data with multiple domains outperforms multilingual data with high in-domain match level and speaker diversity. Unsupervised pretraining does not always outperform the Transformer supervised-only approach, especially when the pretrained data is not diverse enough.
3. Continued pretraining (our best unsupervised pretraining method) is beneficial for both the monolingual and the multilingual scenario regardless of sampling rate mismatch. However, it is more beneficial for less diverse data than the diverse data.
4. We recommend the *Base* architecture instead of $Large_{1\text{-}8}$ for the sake of both accuracy and inference performance. In our experiments Kaiming Initialization shows better performance but the small differences in pretraining might lead to a less clear comparison.
5. The IF Loss works better than the traditional ICE Loss in all data domains. In addition, for the same telephone-domain recognition IF Loss works well but for the large out-of-domain recognition it should only be applied on less diverse pretrained data. In order to further improvement of accuracy, we integrate IF Loss with our prosposed On-off Regularization and L2 Regularization.

Conclusion

Future work

1. Pretraining on the in-domain data (medical) is not compared yet.
2. Data augmentation for finetuning (except SpecAugment) is not investigated yet.
3. Our proposed On-off Regularization for different pretraining schedules is not studied and not compared with sequence discriminative training [Gibson and Hain], which also uses learning rate reset, yet.
4. The novel wav2vec 2.0 - Conformer [Ng et al. 21] has not been investigated on medical domain nor Vietnamese.

Thank you for your attention

Any questions?

References

[Ardila et al. 20] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber.
Common voice: A massively-multilingual speech corpus.
In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4218–4222, 2020.

[Ba et al. 16] J. L. Ba, J. R. Kiros, G. E. Hinton.
Layer normalization, 2016.

[Baevski et al. 20] A. Baevski, Y. Zhou, A. Mohamed, M. Auli.
wav2vec 2.0: A framework for self-supervised learning of speech representations.
In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 12449–12460. Curran Associates, Inc., 2020.

[Beck et al. 19] E. Beck, W. Zhou, R. Schlüter, H. Ney.
Lstm language models for lvcsr in first-pass decoding and lattice-rescoring.
arXiv preprint arXiv:1907.01030, Vol., 2019.

References

[Beulen and Ney 98] K. Beulen, H. Ney.
Automatic question generation for decision tree based state tying.
In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, pp. 805–809, 1998.

[Chen et al. 20] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, Y. Wang.
Data augmentation for children's speech recognition–the" ethiopian" system for the slt 2021 children speech recognition challenge.
arXiv preprint arXiv:2011.04547, Vol., 2020.

[Conneau et al. 20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli.
Unsupervised cross-lingual representation learning for speech recognition.
CoRR, Vol. abs/2006.13979, 2020.

[Doetsch et al. 17] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, H. Ney.
RETURNN: the RWTH extensible training framework for universal recurrent neural networks.

References

In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017.

[Gibson and Hain] M. Gibson, T. Hain.
Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition.
Citeseer.

[Hendrycks and Gimpel 18] D. Hendrycks, K. Gimpel.
Gaussian error linear units (GELUs).
arXiv preprint arXiv:1606.08415, 2018.

[Hsu et al. 21] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, M. Auli.
Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training.
In *Interspeech*, 2021.

[Kneser and Ney 95] R. Kneser, H. Ney.
Improved backing-off for m-gram language modeling.

References

In *1995 International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 181–184 vol.1, 1995.

[Ko et al. 15] T. Ko, V. Peddinti, D. Povey, S. Khudanpur.
Audio augmentation for speech recognition.
In *INTERSPEECH*, 2015.

[Krogh and Hertz 91] A. Krogh, J. Hertz.
A simple weight decay can improve generalization.
In J. Moody, S. Hanson, R. Lippmann, editors, *Advances in Neural Information Processing Systems*, Vol. 4. Morgan-Kaufmann, 1991.

[Luong and Vu 16] H.-T. Luong, H.-Q. Vu.
A non-expert kaldi recipe for vietnamese speech recognition system.
In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pp. 51–55, 2016.

References

[Lüscher et al. 19] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, H. Ney.
RWTH ASR systems for librispeech: Hybrid vs attention - w/o data augmentation.
CoRR, Vol. abs/1905.03072, 2019.

[Lüscher et al. 22] C. Lüscher, M. Zeineldeen, Z. Yang, P. Vieting, K. Le-Duc, W. Wang, R. Schlüter, H. Ney.
Development of hybrid asr systems for low resource medical domain conversational telephone speech.
Baseline systems for Arabic, German, and Vietnamese for HYKIST project. Submitted to ICASSP 2023., Oct. 2022.

[Ng et al. 21] E. G. Ng, C.-C. Chiu, Y. Zhang, W. Chan.
Pushing the Limits of Non-Autoregressive Speech Recognition.
In *Proc. Interspeech 2021*, pp. 3725–3729, 2021.

[Ortmanns et al. 97] S. Ortmanns, H. Ney, X. Aubert.

References

A word graph algorithm for large vocabulary continuous speech recognition.
Computer Speech & Language, Vol. 11, No. 1, pp. 43–72, 1997.

[Ott et al. 19] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, M. Auli.
fairseq: A fast, extensible toolkit for sequence modeling.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, 2019.

[Park et al. 19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le.
Specaugment: A simple data augmentation method for automatic speech recognition.
arXiv preprint arXiv:1904.08779, Vol., 2019.

[Peddinti et al. 15] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, S. Khudanpur.
Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms.

References

2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Vol., pp. 539–546, 2015.

[Schlüter et al. 07] R. Schlüter, I. Bezrukov, H. Wagner, H. Ney.
Gammatone features and feature combination for large vocabulary speech recognition.
In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 649–652, Honolulu, HI, USA, April 2007.

[Sriram et al. 22] A. Sriram, M. Auli, A. Baevski.
Wav2vec-aug: Improved self-supervised training with limited data.
arXiv preprint arXiv:2206.13654, Vol., 2022.

[Tjandra et al. 20] A. Tjandra, C. Liu, F. Zhang, X. Zhang, Y. Wang, G. Synnaeve, S. Nakamura, G. Zweig.
Deja-vu: Double feature presentation and iterated loss in deep transformer networks.

References

In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6899–6903. IEEE, 2020.

[Vaswani et al. 17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin.
Attention is all you need.
Advances in neural information processing systems, Vol. 30, 2017.