Fine-tuning a pre-trained Wav2Vec2 model for Automatic Speech Recognition: experiments on a fieldwork corpus of the Germanic variety from Sauris/Zahre

UNIVERSITÀ DEGLI STUDI DI UDINE

hic sunt futura

Andrea Gulli, Department of Mathematics, Computer Science and Physics
Francesco Costantini, *Emanuela Li Destri*, *Diego Sidraschi*, Department of Humanities and Cultural Heritage

[Laboratorio
[di [comunic@zione
e linguistica]]]

Laboratory of Communication and Linguistics of the Department of Humanities and Cultural Heritage (DIUM) of the University of Udine

GOAL

- Development of semi-automatic transcription methods based on a small amount of annotated data

MOTIVATIONS

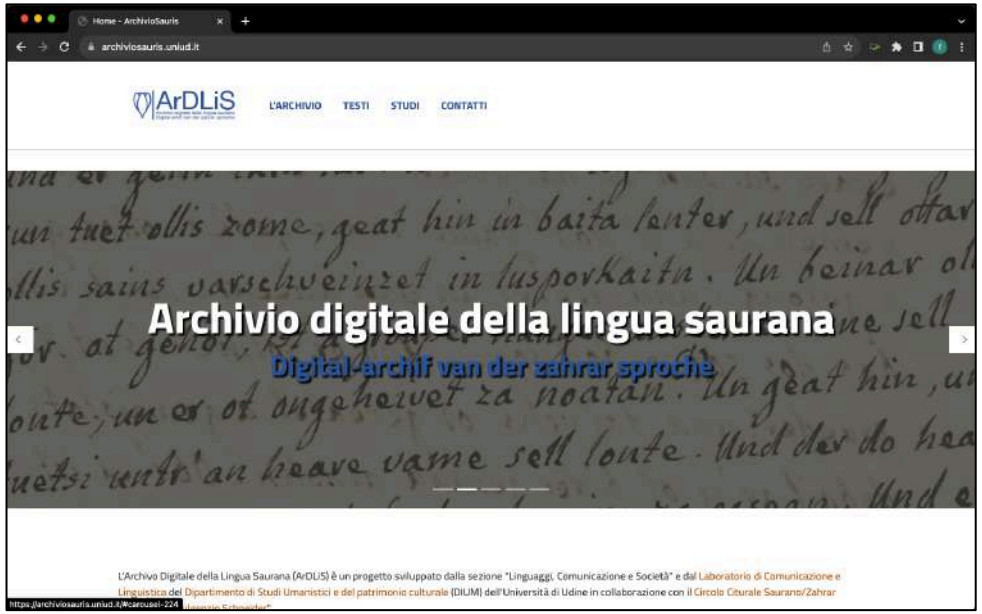
- Protection and valorization of the minority code
- Linguistic research
- Investigation of open questions in the field of automatic speech recognition (ASR)
    - Evaluation of the minimum necessary size of a corpus for training
    - Hyperparameters tuning effect on different linguistic corpora

Sauris and Saurian

- Three villages: Plozn (Sauris di Sopra), Dörf (Sauris di Sotto), Lateis
- ≈400 inhabitants
  - “Triglossia” (Denison 1968)
    - Italian: H-variety
    - Friulian: M-variety
    - Saurian: L-variety
  - Denison (1992):
    - Italian has extended the domains of use to the detriment of Friulian and Saurian
    - Italian is essentially the exclusive code in the younger generations
- Nowadays ≈200 speakers of Saurian (Costantini 2021)
  - sharp decrease in the number of speakers

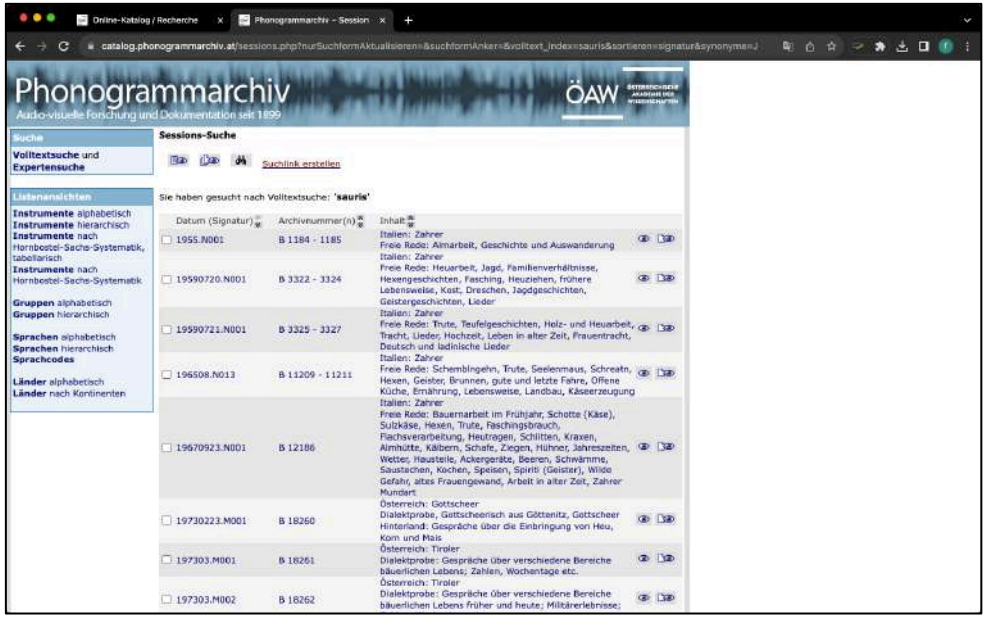
Sauris and Saurian

- Need of language documentation for Saurian
  - Collecting, describing, and archiving linguistic data
  - ArDLiS (digital archive collecting texts in Saurian, 1800-)



Sauris and Saurian

- Recordings preserved at the Phonogrammarchiv of the Austrian Academy of Sciences (ÖAW)
  - 21 recordings
  - total duration of around 8 hours
  - interviews with speakers from Sauris
  - collected between 1955 and 1986
    - Eberhard Kranzmayer, Maria Hornung
    - autobiographies, customs, and traditions



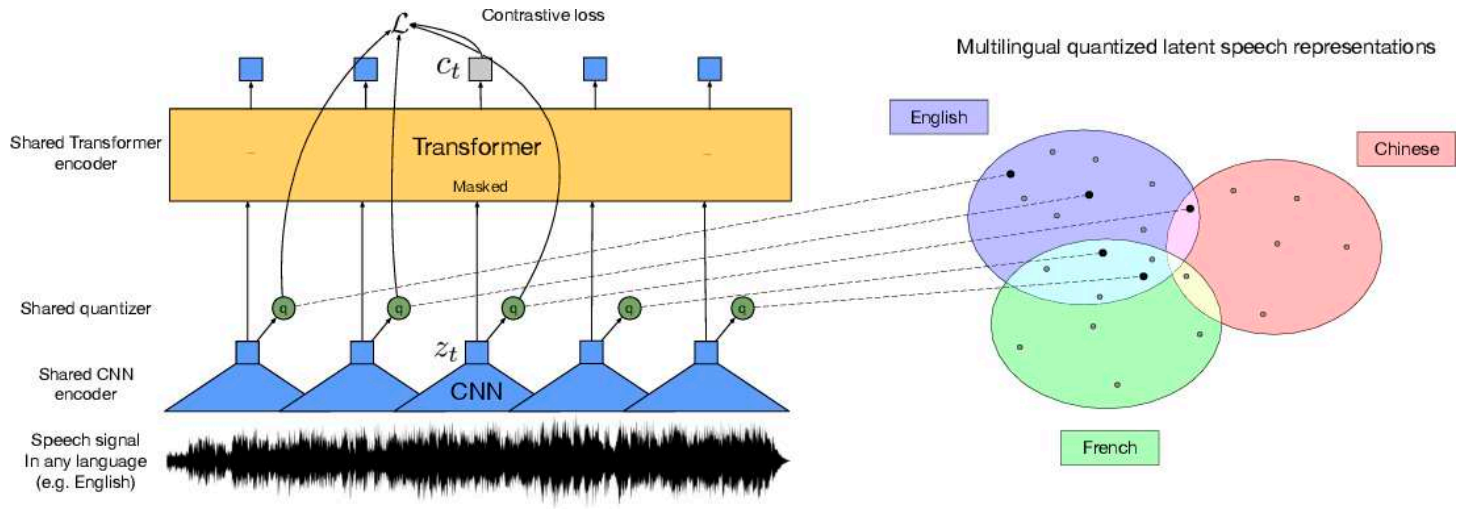
New collected data

- Recordings made in Sauris in 2023
  - 6 Saurian language speakers, 46 recordings
  - non-spontaneous speech: “Le nostre parole - Unsere Wörter”, Umberto Patuzzi
  - not wordlists, but small sentences with complete meaning (simple utterances, songs, nursery rhymes)
  - ≈1.5 hours in total
  - WAV files, 44.1 kHz, Stereo
  - Utterances of ≈1.5 seconds
  - Common Voice structure:
    - path to each audio sentence
    - sentence of each audio file
    - speaker of each sentence

Model

XLSR-Wav2Vec2 learns powerful speech representations from hundreds of thousands of hours of speech in more than 50 languages of unlabeled speech (*XLSR: cross-lingual speech representations*)

- learning contextualized speech representations by randomly masking feature vectors before passing them to a transformer network
- language-specific fine-tuning on very little labeled data



Fine-tuning

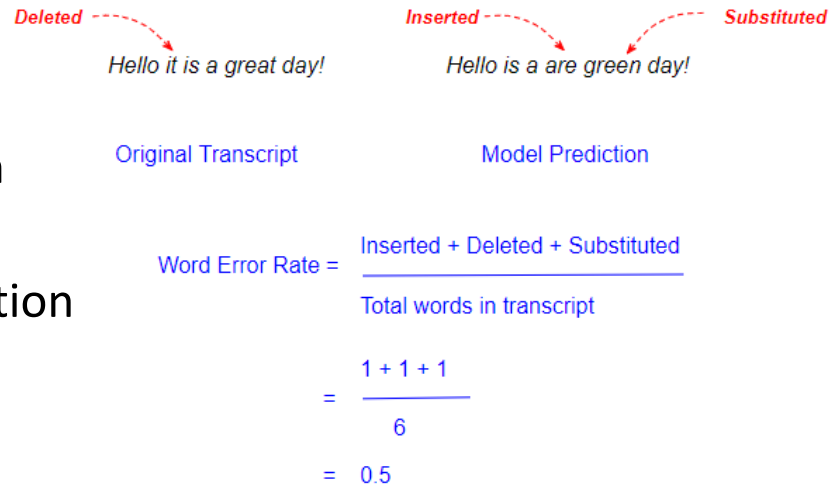
- ASR models transcribe speech-to-text
  - feature extractor that processes the speech signal to the model's input format
  - tokenizer that processes the model's output format to text
- Fine-tuned models map the sequence of context representations to its corresponding transcription
  - linear layer on top of the transformer block classifying context representation to token class: output size is the number of tokens in the *vocabulary*
  - fine-tuning with *Connectionist Temporal Classification (CTC)*: algorithm used to train neural networks for sequence-to-sequence problems

Preprocessing

- Split the dataset in training (80%), validation (10%), and testing (10%) sets
- String preprocessing
  - Remove punctuation symbols: , .:;?!'"*«»“ ”—-‘’…
  - All lowercase characters
  - Extract characters of each sentence
- Create the vocabulary
  - Join all the sentences in one big sentence and create the union of all distinct characters
  - Convert the resulting list into an enumerated alphabet
  - " " has its own token class: |
  - "unknown" token: to deal with characters not encountered in the training set
  - "padding" token: corresponds to CTC's "blank token”
- Audio preprocessing
  - mp3 conversion
  - downsampled to 16 kHz
  - Mono
  - Zero-mean-unit-variance normalized

Loss and Evaluation metrics

- CTC loss
  - The sum of all valid alignments between the predicted and target sequences
  - Negative logarithm of the probability that the predicted sequence aligns with the target sequences
  - Mean (average) across all time steps and items of the batch
- Word Error Rate
  - The most used metric for Speech-to-Text problems
  - Comparing prediction and target word-by-word
  - Deletion: present in the transcription, missing in the prediction
  - Insertion: not in the transcription, added in the prediction
  - Substitution: altered between the prediction and the transcription
  - Human transcribers: WER ≈4%.
  - Commercially available ASR software (in English!): WER ≈12%



- Character Error Rate
  - Comparing the predicted output and the target transcript character-by-character

Training

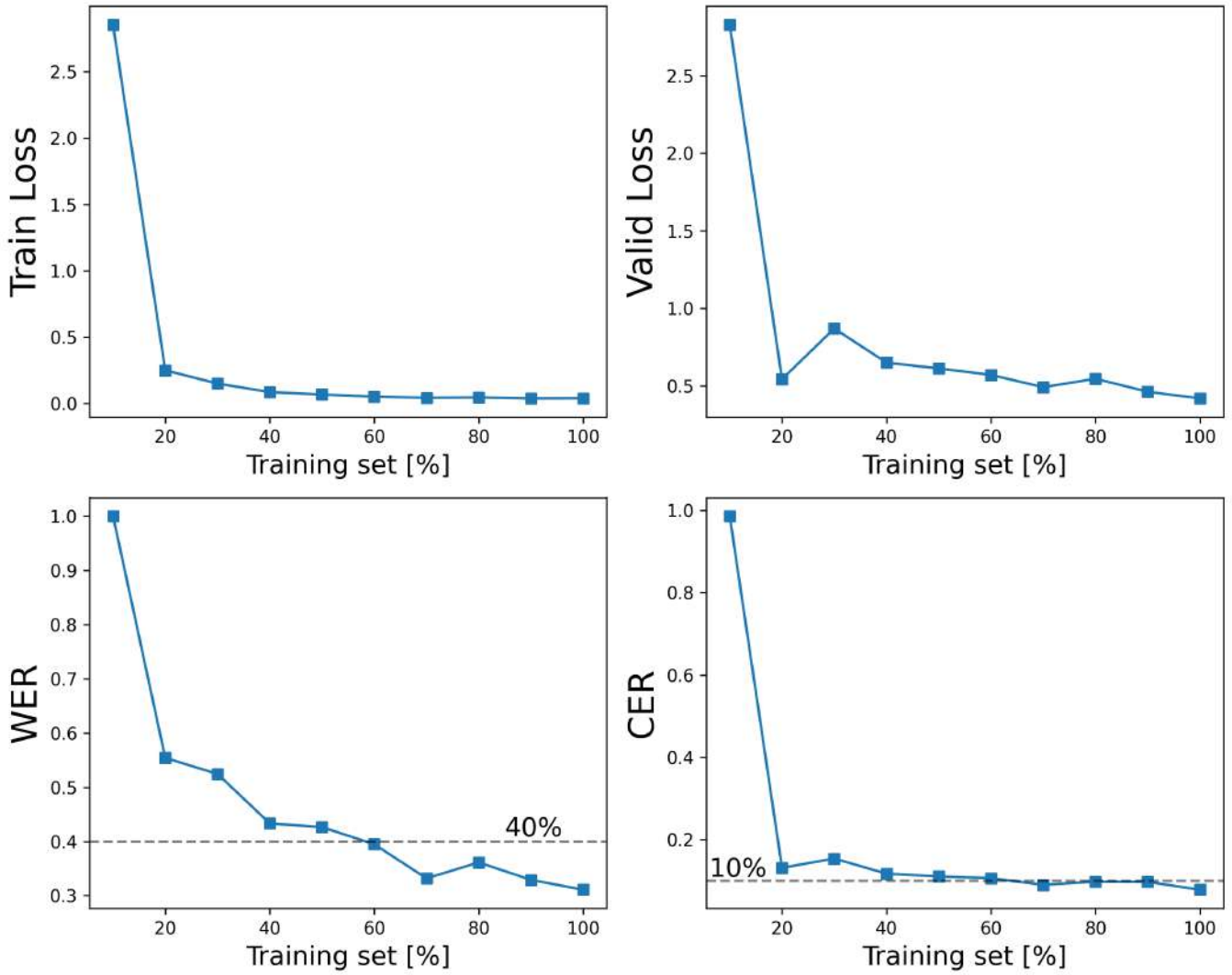
- Data collator
  - Pad the training batches dynamically to the longest sample in their batch and not the overall longest sample
  - Separate padding for inputs and labels
  - Predictions generated on the padding tokens are not taken into account when computing the loss
- Load and configure pre-trained checkpoint
- Define the training configuration
  - Group training samples of similar input length into one batch
  - Learning rate (3e-4) heuristically tuned until fine-tuning has become stable
- Training specifics
  - 60 epochs
  - Evaluate and log metrics every 50 steps, save every 100 steps
  - Dropout regularization parameters:
    - *Attention dropout*: randomly drops some attention weights within the transformer architecture (0.1)
    - *Hidden dropout*: randomly deactivates a fraction of neurons within each hidden layer including the transformer's encoder and decoder (0.1)
    - *Masked dropout*: specifies the probability of masking each time step in the input sequence (0.075)
    - *Layer dropout*: specifies the probability of dropping out an entire hidden layer (0.1)
- ≈2h on a Google Colab Tesla T4 TPU

Results: best performing model

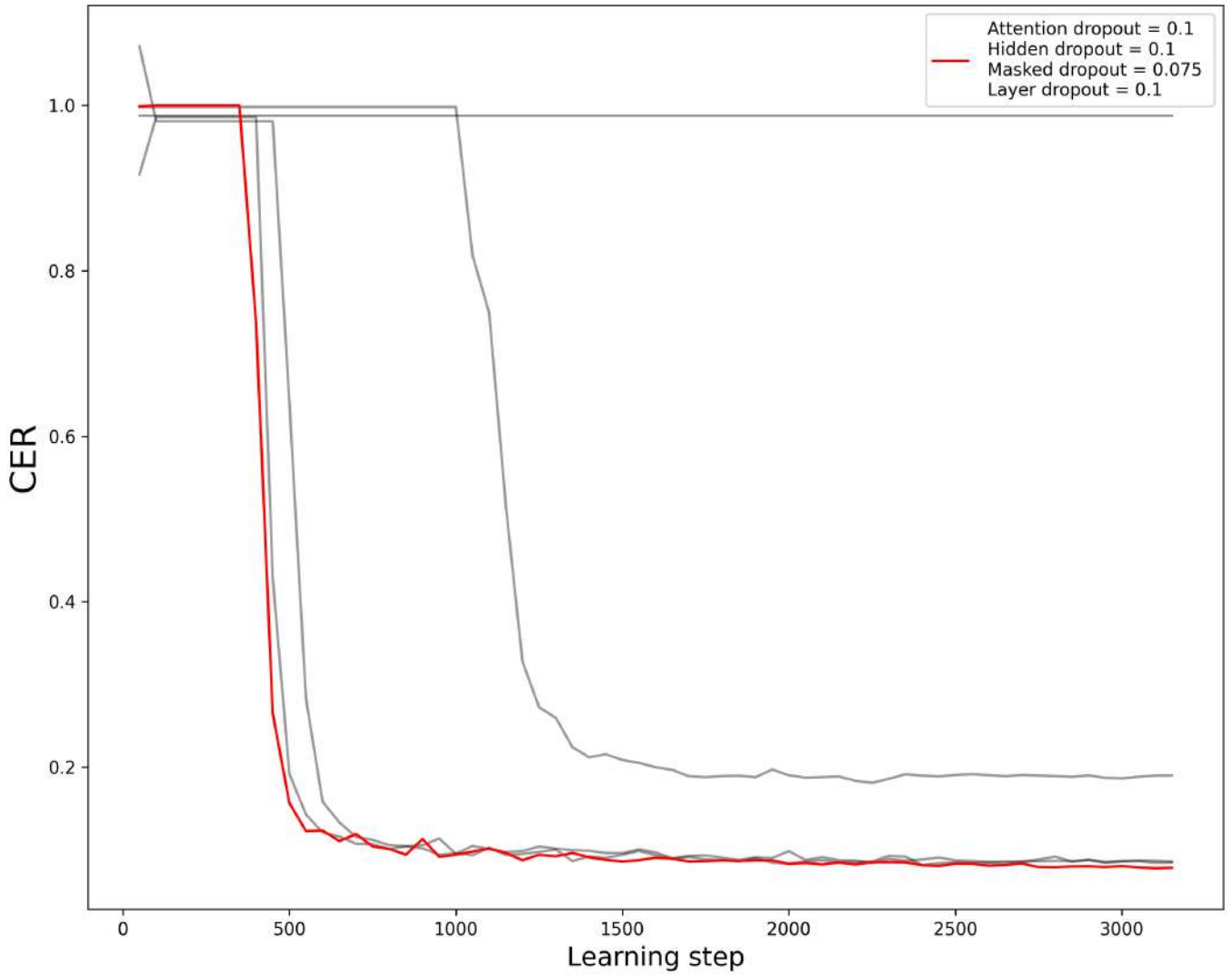
- Metrics
  - Test set of ≈100 utterances (length: $\mu = 8.47$, $\sigma = 2.78$)
  - CER: $\mu = 5.33\%$, $\sigma = 7.33\%$
  - WER: $\mu = 21.89\%$, $\sigma = 22.30\%$
- Examples

Reference	Prediction	WER	CER
ober ben as ar ist aufgestean sentse schon geben gemochater	ober ben as ar ist aufgestean sentse schon geben gemochater	0.0	0.0
im bolde seint geben abesn scheana pliemblan	im bolde seint geben abesn scheana pliemblan	0.0	0.0
as der on gemusset intrinen	as as on gemusset in trinen	0.6	0.14814814814814814
d ot ois gesot ana schia filastroka	d otis ois gesot ana schia filastrocka	0.2857142857142857	0.11428571428571428
de seint börtn gesaht ame longas sel ime maie	de seint börtn gesaht ame longas sel ime maie	0.0	0.0
unt seint glaich gean za mahn	unt seint glaich gean za mahn	0.0	0.0
bie hasseste i hasse luigi	bie hasseste i hasse luingi	0.2	0.038461538461538464
tschnos ontse gevairt ola minonder	schnos ontse gevairt ola minonder	0.2	0.029411764705882353
s ist böl oise dot gemusset mochn tschölschoft in der oltn muma	s ist ber olse d ot gemusst mochn tscheischoft in deroltn muma	0.6666666666666666	0.12698412698412698
unt sent a khemen khaufars as dontme gebn gelt	unt seint akhemen khaufars as dontme gebn gelt	0.3333333333333333	0.043478260869565216

Results: training set size



Results: hyperparameters tuning



Ongoing work and open issues

- Language model (LM)
  - Integration with a KenLM
  - Which language for the LM?
  - Which dataset for the LM?
- Next challenges
  - Punctuation
  - Phonogrammarchiv noisy data (multiple overlapping speakers, heavy background noise)
- Hyperparameters grid search
  - Larger ranges
  - More hyperparameters
- Present and share the results
  - multimedia page on the ArDLiS website
  - sharing through endangered languages repositories (Pangloss, Common Voice)