DATASET

Common Voice is a publicly available voice dataset, powered by the voices of volunteer contributors around the world.

Data used: **Common Voice Delta Segment**, contains clips of variable length

	Version	Number of speakers	Recorded hours
Spanish	12.0	373	57
Japanese	14.0	77	54
Italian	12.0	65	13

Expectation: Spanish and Italian to perform similarly, Japanese to perform very differently.

PREPROCESSING

Common voice Delta segments audio clips

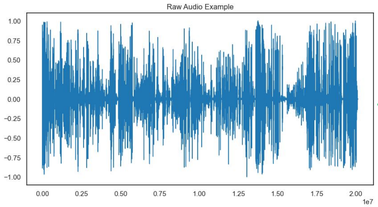
↓

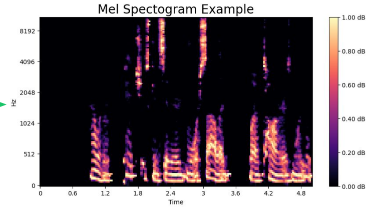
5 seconds from start of clip

if empty, add silence

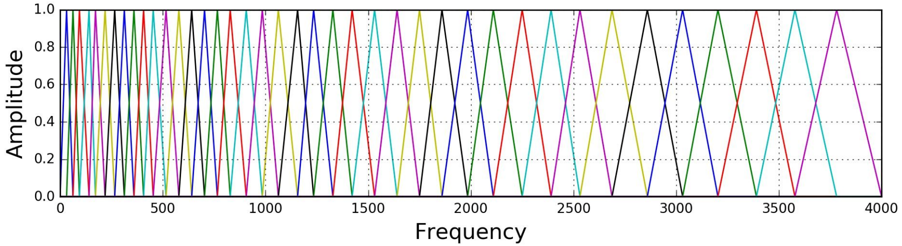
↓

Convert to Mel spectrograms:
visual representation of the frequency content of an audio signal.



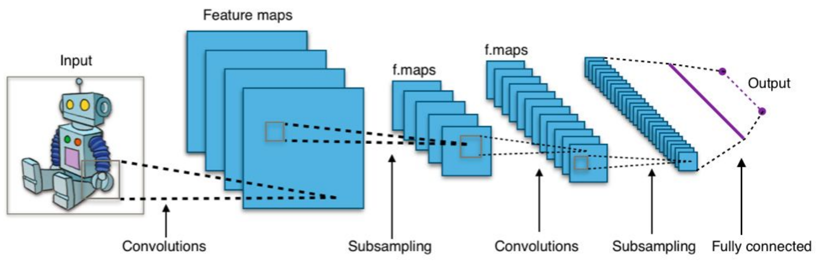


1. Audio signal→short overlapping frames
2. Fourier Transform to obtain the frequency spectrum
3. Mel-frequency bins are used to align with human perception.
4. The mel filterbank to convert the linear frequency scale to the mel scale.

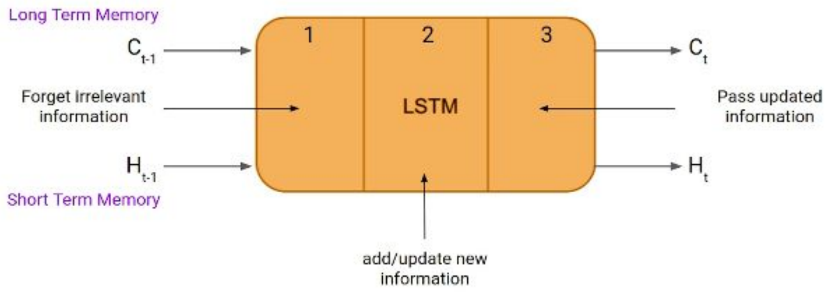


MODELS

CNN: Convolutional Neural Network



LSTM: Long Short Term Memory



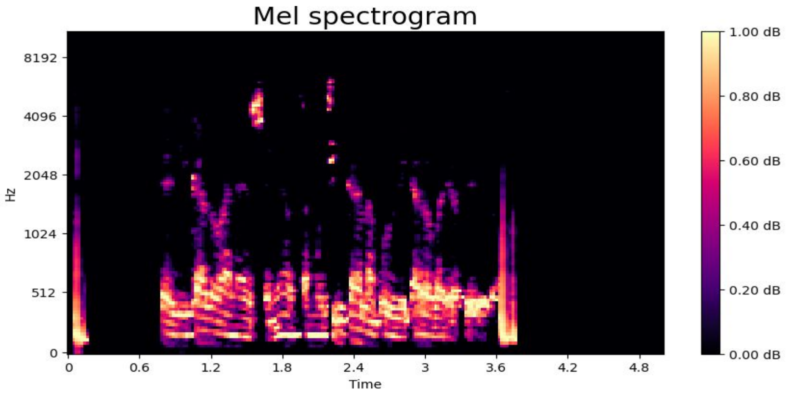
Sequence to sequence:

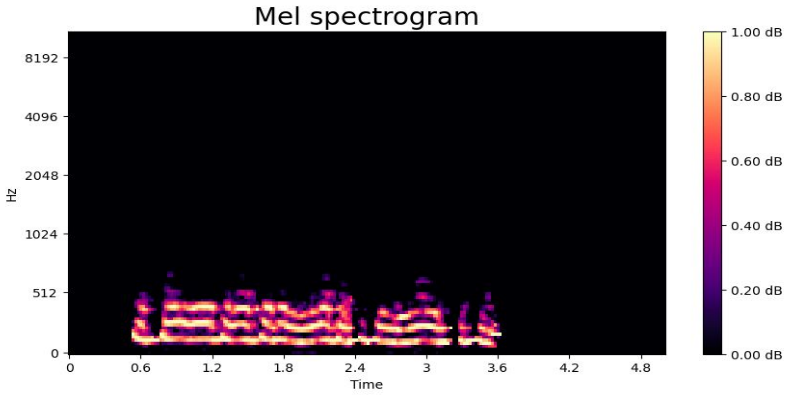
- ❑ Regressor predicting a time window
- ❑ Regressor predicting next time step

EXTRACT and ANALYSE **ENCODED SPACE**

TRAINING PROCEDURE

1. Train model with Spanish (chosen as "native" language).
2. - Pretrain the model with Spanish, Japanese or Italian **low pass filtered** (500 Hz), with a subset of the data;
   - Continue training with the whole dataset of Spanish.





PRE-TRAINING EFFECT ON TIME

We'll have a look of how:

- ❑ No pretraining
- ❑ Pre-training with same language as training
- ❑ Pre-training with different languages than training

affects the **time** to reach a certain **threshold** in the loss value.

PRE-TRAINING EFFECT ON LATENT SPACE

3 languages equally distributed

↓

Trained models

↓

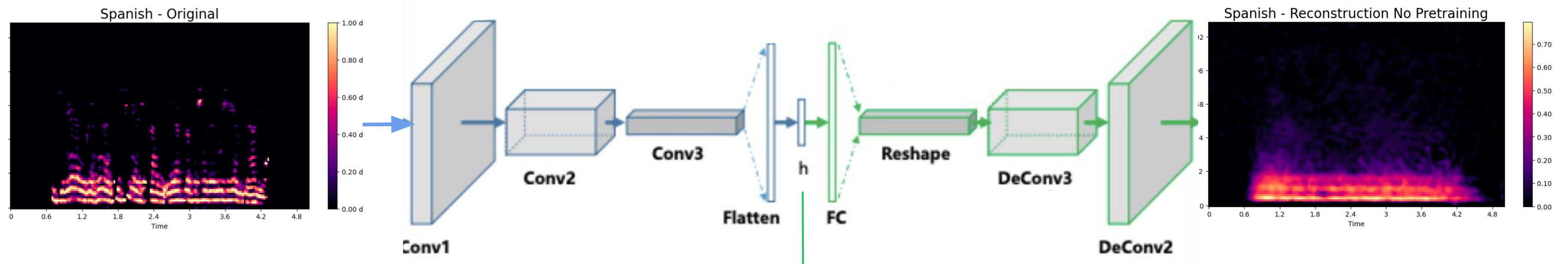
Latent space ⇆ structure of internal representation

↙ ↘

Binary classifier

t-SNE

CNN AUTOENCODER



Encoder

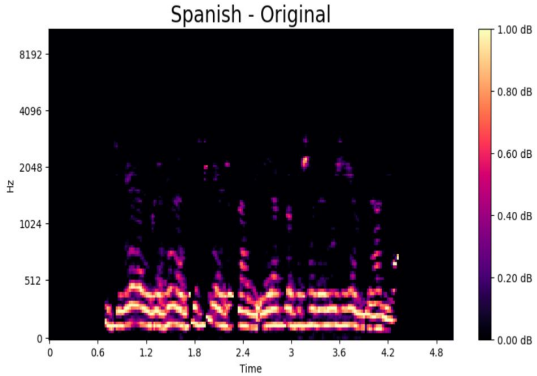
Decoder

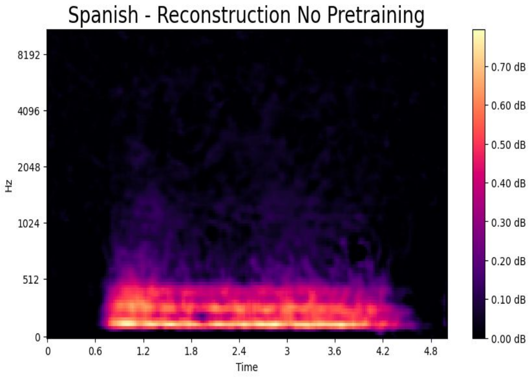
Latent Space:

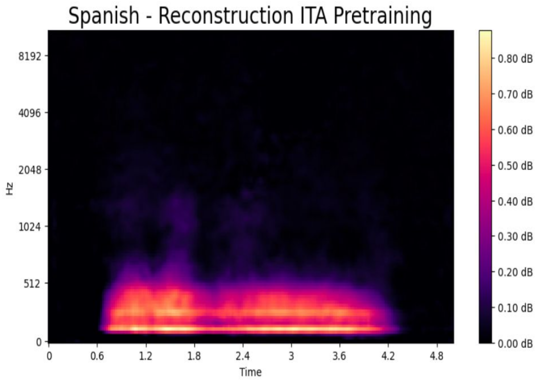
Linear layer

3000 neurons

CNN AUTOENCODER: Results







Best loss reached within all configurations:
0.016

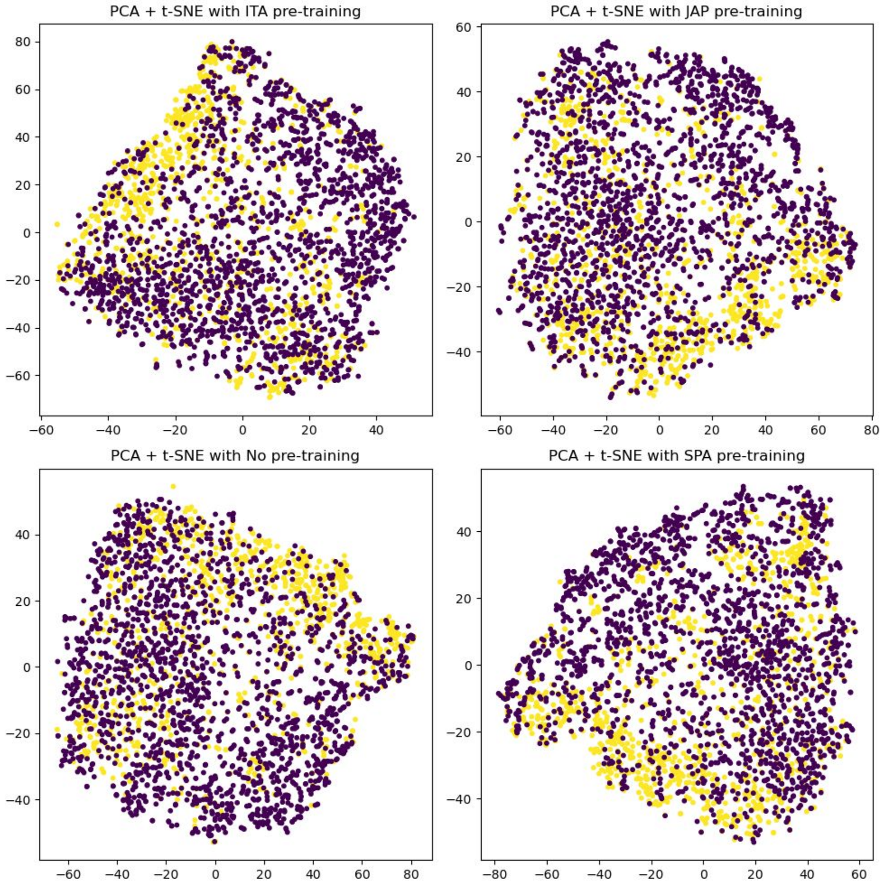
CNN AUTOENCODER: Results

Is NATIVE language **distinguishable** from NON-NATIVE in the 3000dim - Latent Space?

Pre-training	Spanish	Italian	Japanese	None
Perceptron accuracy	87.1 %	87.2 %	87.1%	87.1 %

CLASSIFICATION TASK ✓

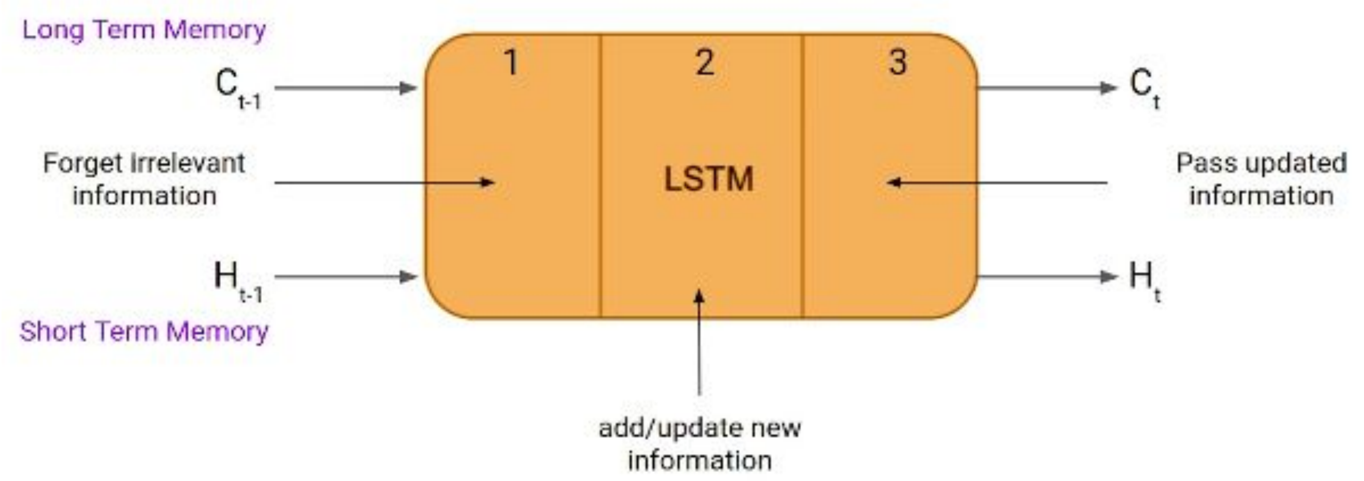
CNN AUTOENCODER: Results



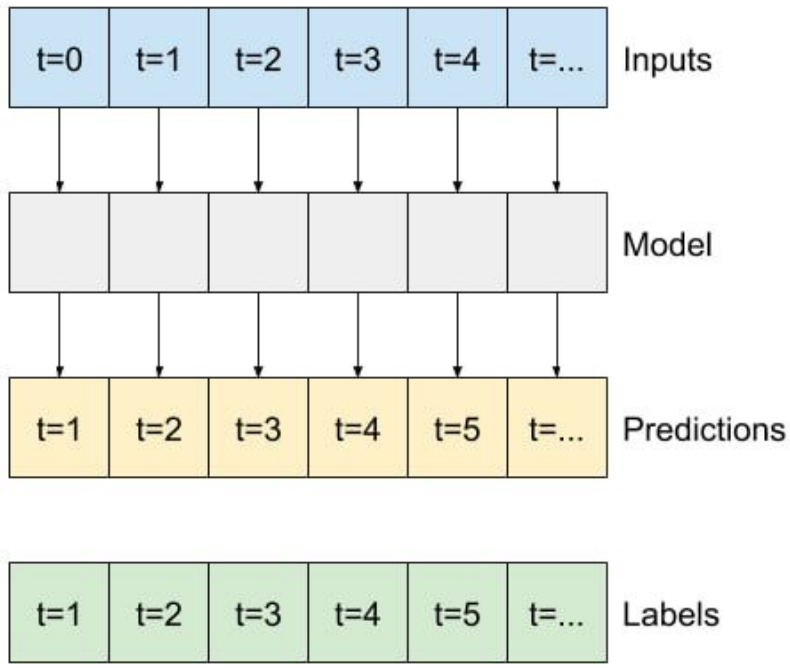
t-SNE ✖

LSTM: Long Short Term Memory

LSTM: Type of **recurrent neural network** (RNN) architecture designed to **handle long-term dependencies** in sequential data.



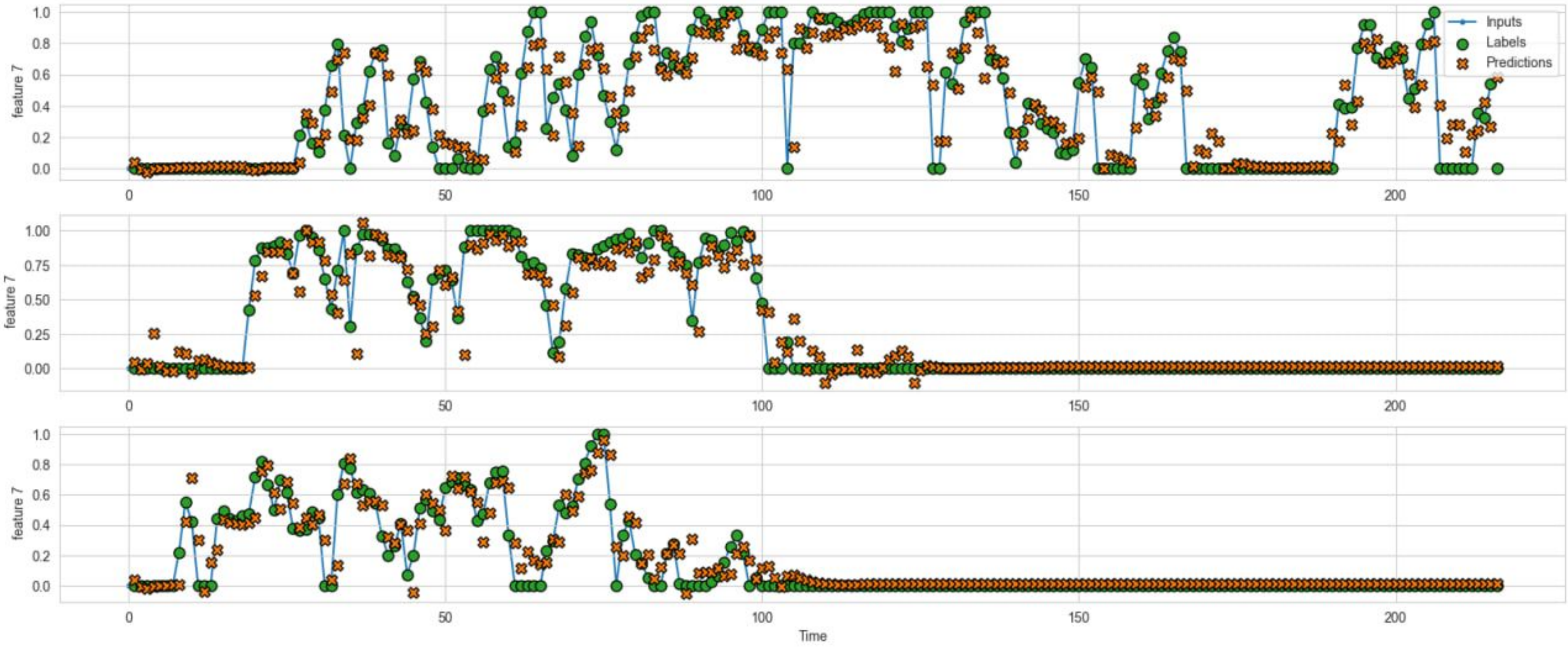
LSTM Sequence to sequence multivariate regressor



- The LSTM will accept as input a series of time steps decompositions of a spectrogram and will adjust its weights in order to predict the next time step.
- The model given t predicts t+1

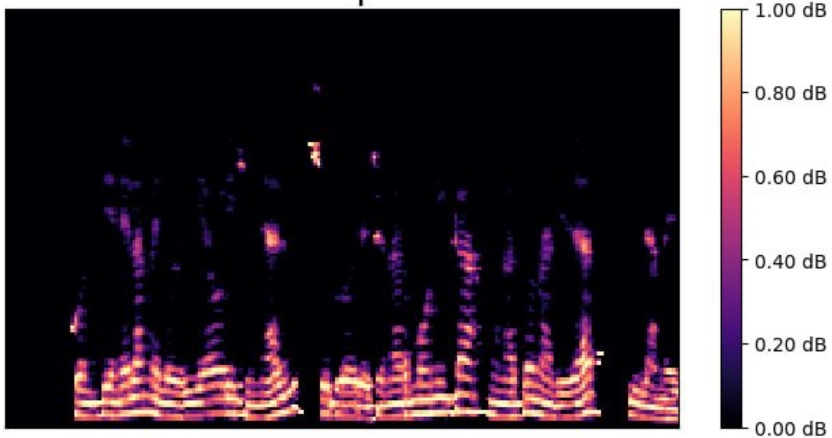
```
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 216, 128)]        0

 lstm (LSTM)                 [(None, 216, 250),        379000
                              (None, 250),
                              (None, 250)]

 dense (Dense)               (None, 216, 128)          32128

=================================================================
Total params: 411128 (1.57 MB)
Trainable params: 411128 (1.57 MB)
Non-trainable params: 0 (0.00 Byte)
```

LSTM Sequence to sequence multivariate regressor

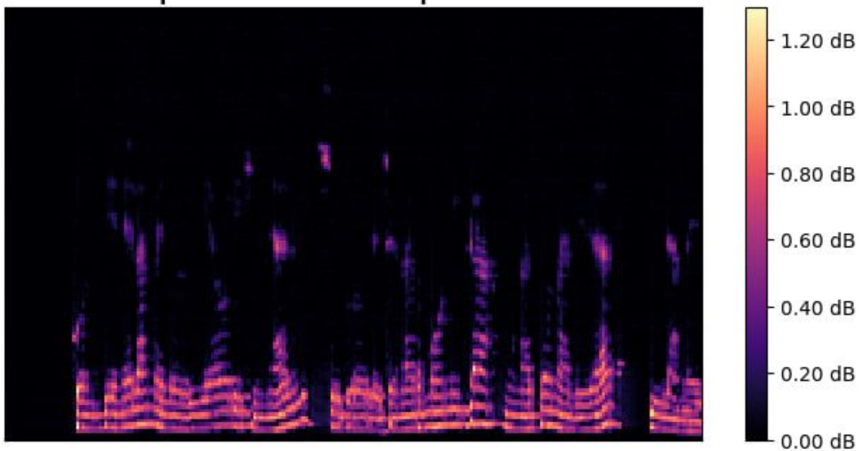


LSTM Sequence to sequence multivariate regressor

Real Sequence

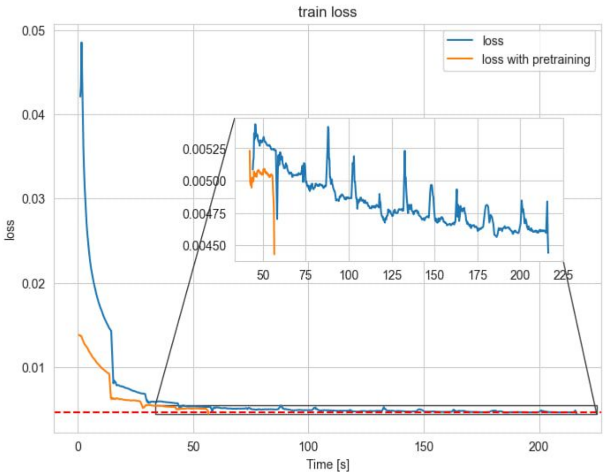


predicted Sequence

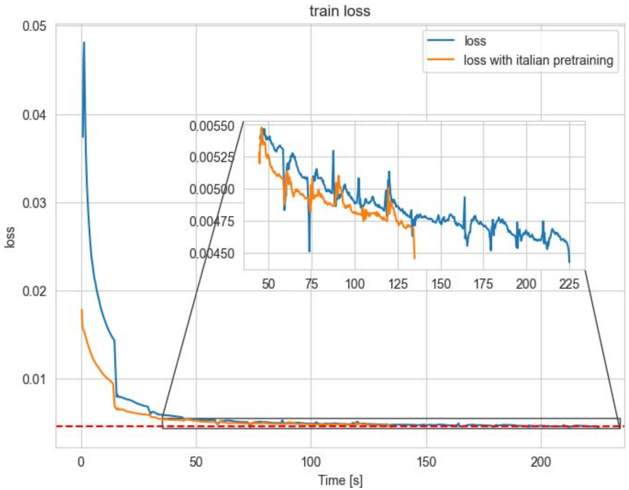


LSTM Sequence to sequence multivariate regressor

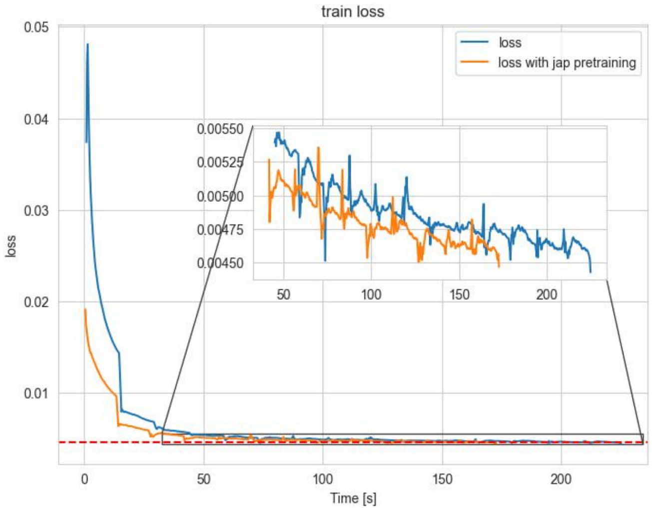
Spanish pre-train



Italian pre-train



Japanese pre-train



Threshold: 0.0044

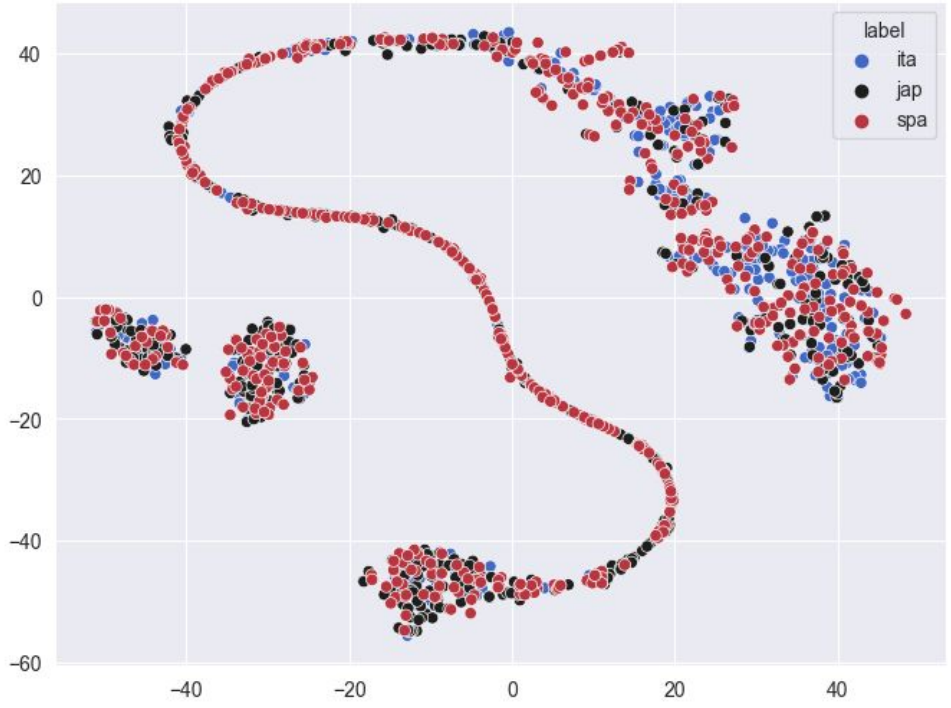
LSTM Sequence to sequence multivariate regressor

SVM accuracy on last hidden state and cell state

	No pre-train	Spanish pre-train	Italian pre-train	Japanese pre-train
Cell	60.6%	62.8%	72.0%	61.9%
Hidden	60.9%	63.2%	71.7%	60.4%

LSTM Sequence to sequence multivariate regressor

T-sne on cell states



No meaningful results in the encoded space