Progress Report: Week I

Esperanto

Exchanges for SPEech ReseArch aNd TechnOlogies

Horizon 2020 project

Overview

- Brief review of our plans and goals
- How we divided our work
- Encoding Team progress report
- Decoding Team progress report
- Analysis Team progress report
- Q&A

Plans and goals

- Develop a Multi-Modal / Multi-Lingual / Extensible Translation system
  - Text/Speech inputs and outputs
  - Assume common multi-lingual space
  - Easily add new languages with low resources

Work division

- Three working groups with strong interrelation
  - Encoding
  - Decoding
  - Analysis

Encoding Team
Common representation space

- **1st goal:** Learn a multilingual semantically aligned embedding space for speech and text
  - For text, already have LaBSE[1] and LASER[2]
  - Need something similar for speech: XLS-R[3] is multilingual, but does not project semantically aligned sentences in the same space
  - Starting point: **SAMU-XLSR**[4]

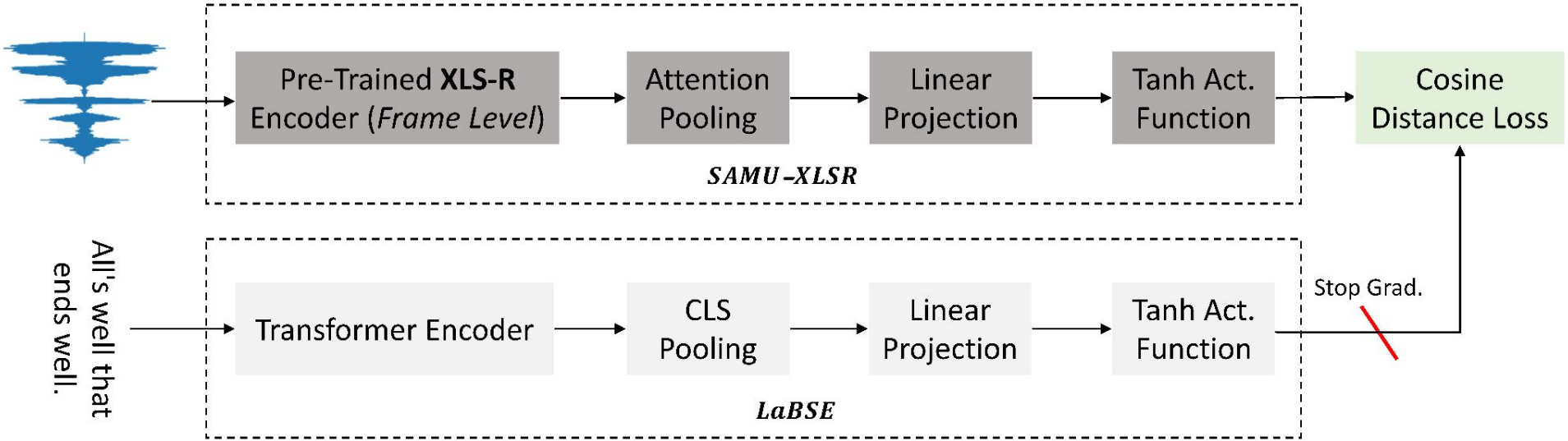
[1] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic BERT Sentence Embedding,” 2020. [Online]. Available: https://arxiv.org/abs/2007.01852
[2] H. Schwenk and M. Douze, “Learning joint multilingual sentence representations with neural machine translation,” 2017. [Online]. Available: https://arxiv.org/abs/1704.04154
[3] A. Babu et al., “ XLS-R: Self-supervised Cross-lingual Speech Representation at scale,” 2018. [Online]. Available: https://arxiv.org/abs/2111.09296
[4] S. Khurana, A. Laurent, J. Glass, “SAMU-XLSR: Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation”, 2022. [Online] Available: https://arxiv.org/abs/2205.08180

Encoding Team
SAMU-XLSR

- Multimodal embedding vector space (text & speech)
- Cross-lingual (trained on 51 languages from CommonVoice-v8)
- And semantically aligned (Wav2vec 2.0 and XLS-R are not)
- Fine-tune of pre-trained XLS-R via knowledge distillation from LaBSE
- + Pooling mechanism and non-linear projection layer

⇒ Transform the frame-level contextual representations into a single utterance level embedding vector

Encoding Team
SAMU-XLSR (training)

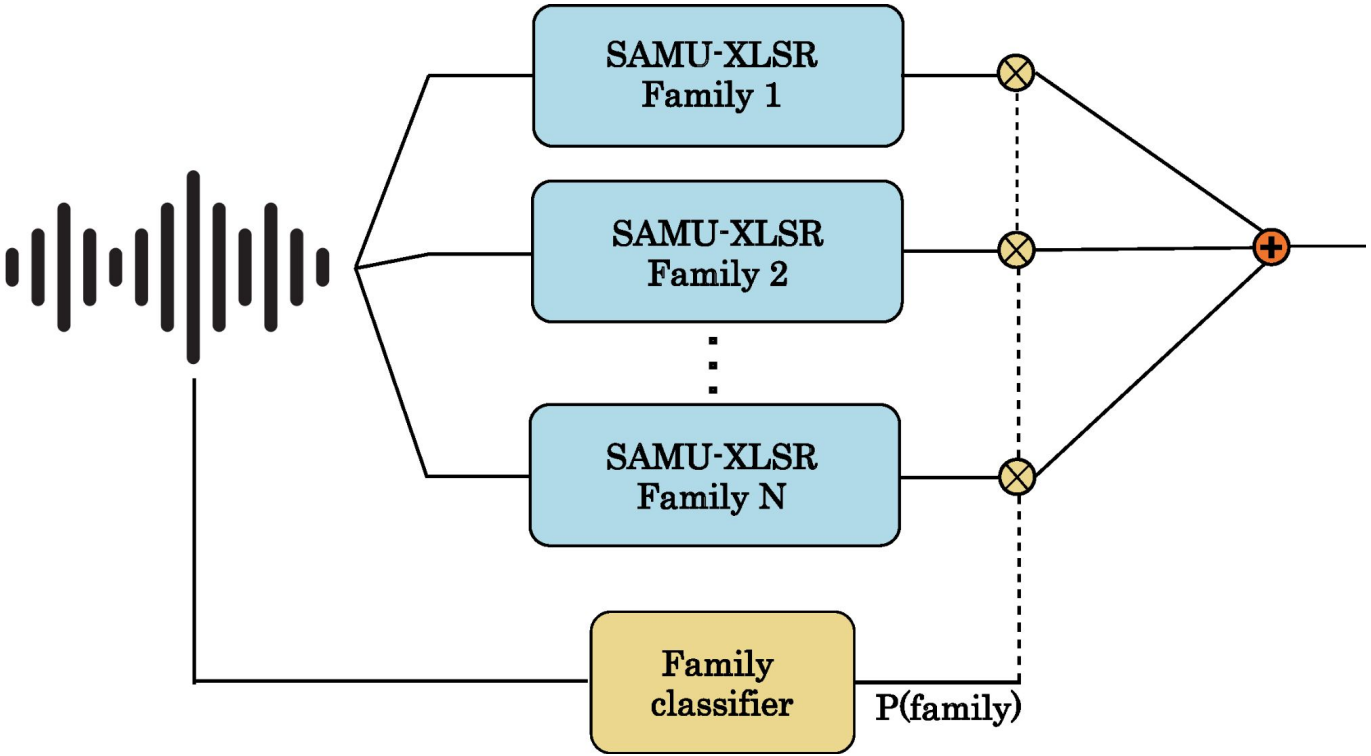


S. Khurana, A. Laurent, J. Glass, 2022. [Online] Available: https://arxiv.org/abs/2205.08180

Encoding Team
Expand for low-resource languages

- **2nd goal:** Be able to use the system with unseen low-resource languages
  - Fusion of language-family based models for speech
  - New multilingual pre-trained LM for low-resource MT / ST for text

Encoding Team
Fusion of speech models



Encoding Team
Fusion of speech models

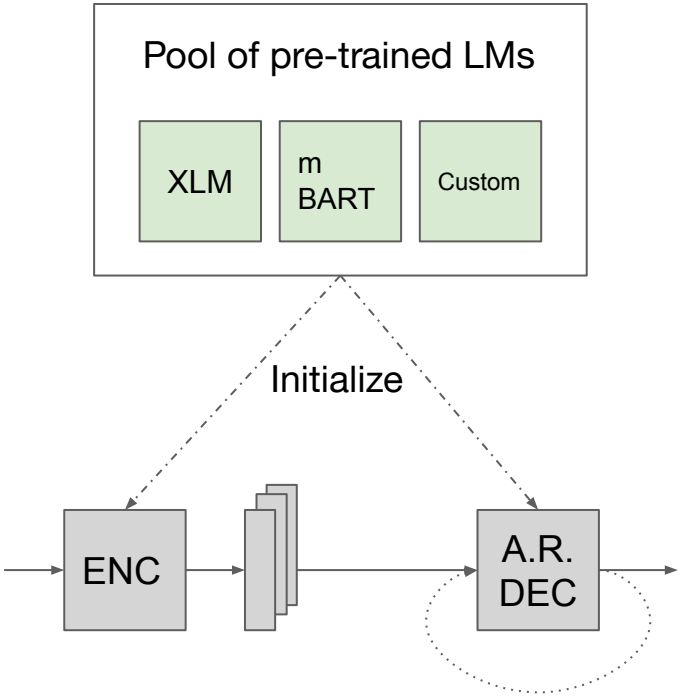
- One embedding (or sequence of embeddings) per language family
- + Fusion of models during inference, based on the scores of an independent language-family ID system: weighted average of embeddings
- Different strategies for training the language-family embeddings
  - One SAMU-XLSR per language family
  - Only one XLS-R (fine-tuned for SAMU-XLSR) module + pooling mechanism and non-linear projection layer trained separately for each language family

Encoding Team
Fusion of speech models

- Language families:
  - Italic: ca, fr, es, it, pt, ro, gl
  - Germanic: en, de, nl, sv, da
  - Balto-Slavic: be, ru, pl, uk, cs, lt, sk, bg, sl, lv
  - Indo-Iranian: fa, hi
  - Chinese: zh-HK, zn-CN
  - Volta-Congo: rw, sw
  - Berber: kab
- Fine tune with 150 h from Common Voice 8.0 for each family

Encoding Team
Pre-training LMs for MT / ST

- Pre-trained LM can be used to initialize encoder and/or decoder in a standard MT / ST pipeline.
- Empirically studying the effect of various pre-training objectives / models for downstream MT / ST.
- Working on a new seq2seq architecture for LM (pre-)training.
- Additionally making use of bilingual dictionaries for low-resource languages.



Encoding Team
Data Collection

- We had some text sources in Tamasheq/Tamahaq
  - A language learner's book (PDF) and 2 Bible translations (Text Files)
- We ended with the two Bible translations, and about 1000 lines of Tamasheq to English translations in two main dialects — Tamaghit and Tudalt
  - The Bibles are monolingual
  - The 1000 lines are translations of various words and sentences
- We also used the HOW2 and MuST-C datasets to simulate low-resource languages with various models

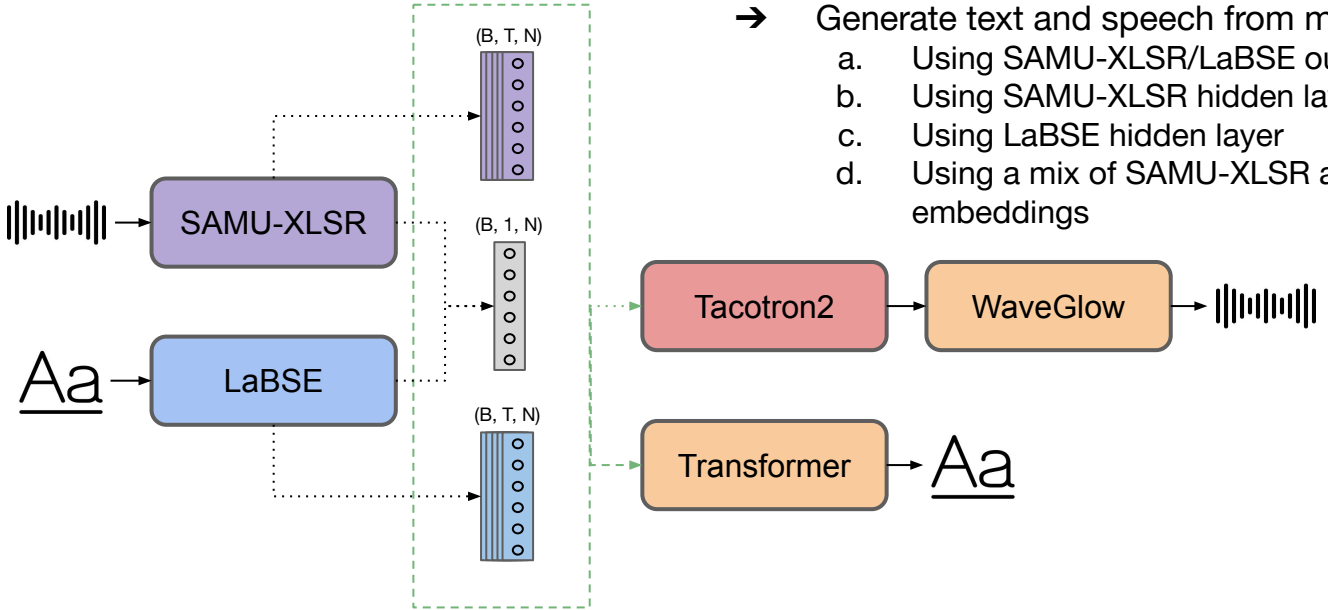
Encoding Team
Model Experimentation

- First, we wanted to simulate low-resource languages
  - HOW2 was already prepped to have several splits
- We made a script that could split the MuST-C datasets into various splits for low-resource simulation
- We are currently still fine-tuning various models to see which ones perform the best on low-resource languages for textual data

Decoding Team
Goal

- Generate text and speech from the common space
- Evaluate audio-only outputs (speech2speech metrics)

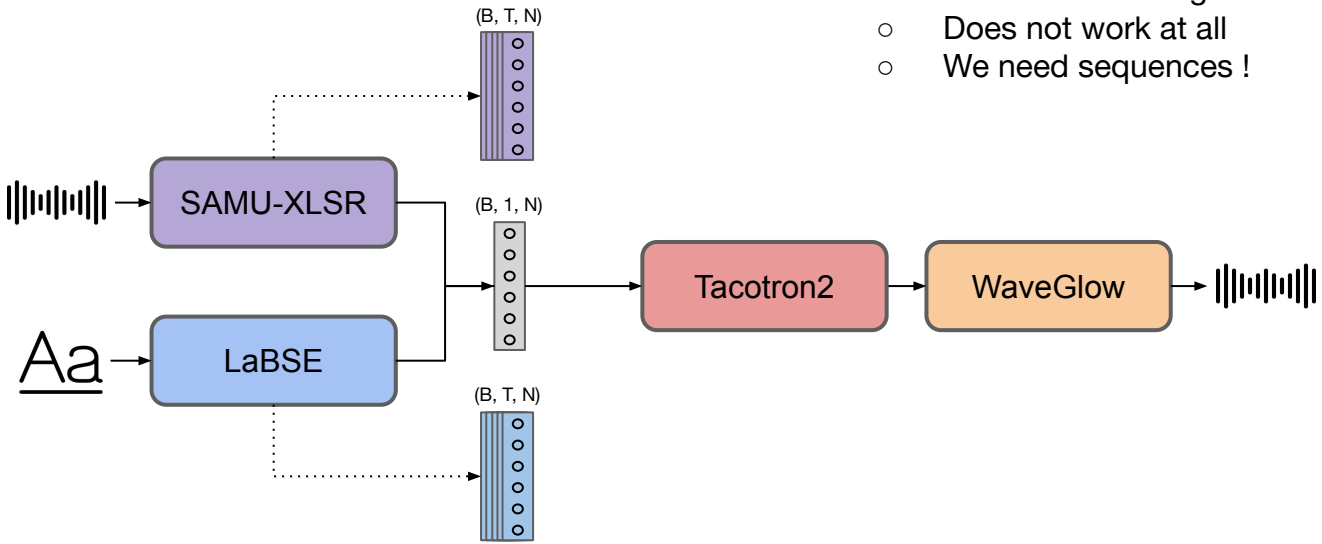
Decoding Team
SAMU-XLSR & LaBSE



➔ Generate text and speech from multilingual space
- a. Using SAMU-XLSR/LaBSE output
- b. Using SAMU-XLSR hidden layer
- c. Using LaBSE hidden layer
- d. Using a mix of SAMU-XLSR and LabSE embeddings

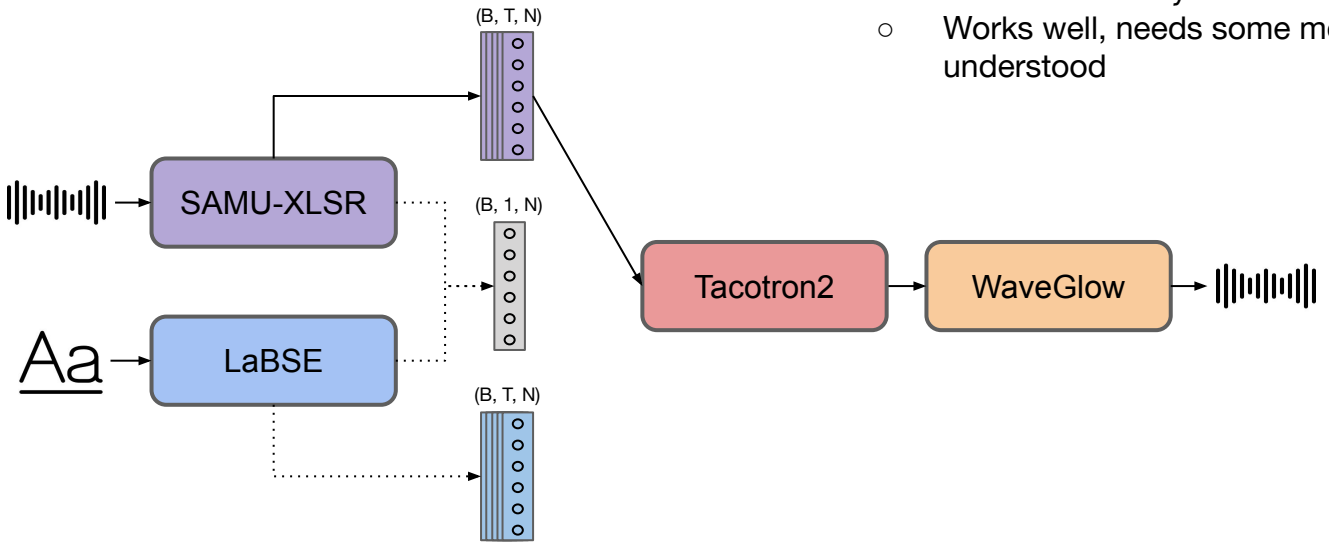
[1] A. Babu et al., " XLS-R: Self-supervised Cross-lingual Speech Representation at scale," 2018. [Online]. Available: https://arxiv.org/abs/2111.09296
[2] S. Khurana, A. Laurent, J. Glass, "SAMU-XLSR: Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation", 2022. [Online] Available: https://arxiv.org/abs/2205.08180

Decoding Team
SAMU-XLSR & LaBSE



➔ SAMU-XLSR/LaBSE single embedding
- Does not work at all
- We need sequences !

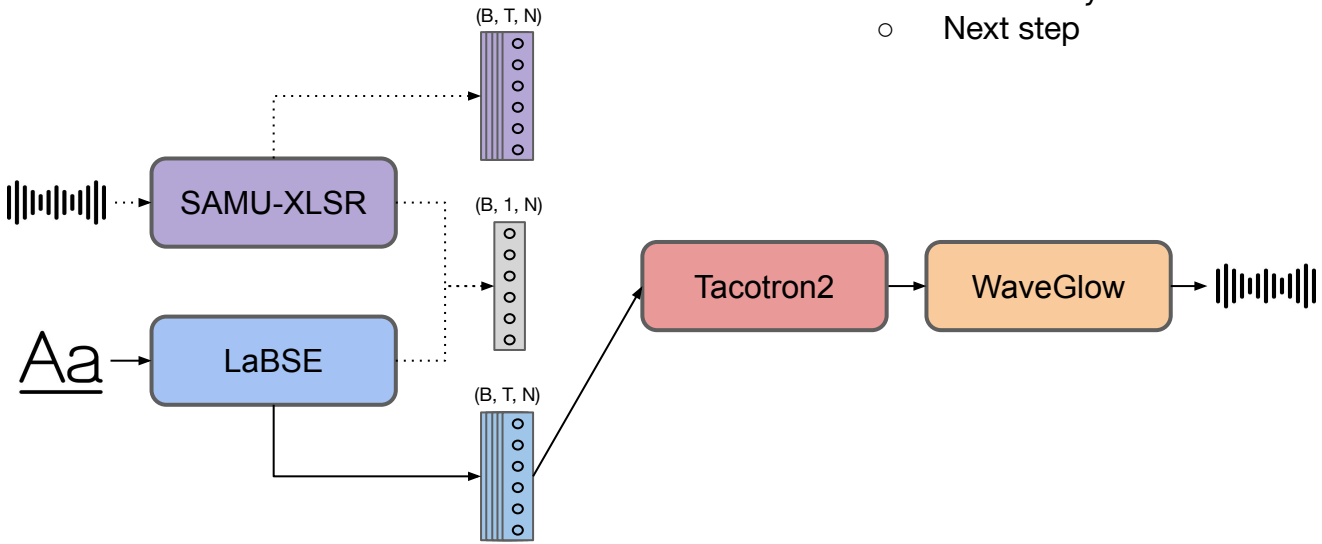
Decoding Team
SAMU-XLSR & LaBSE



➔ SAMU-XLSR hidden layer
- Works well, needs some more training to be understood

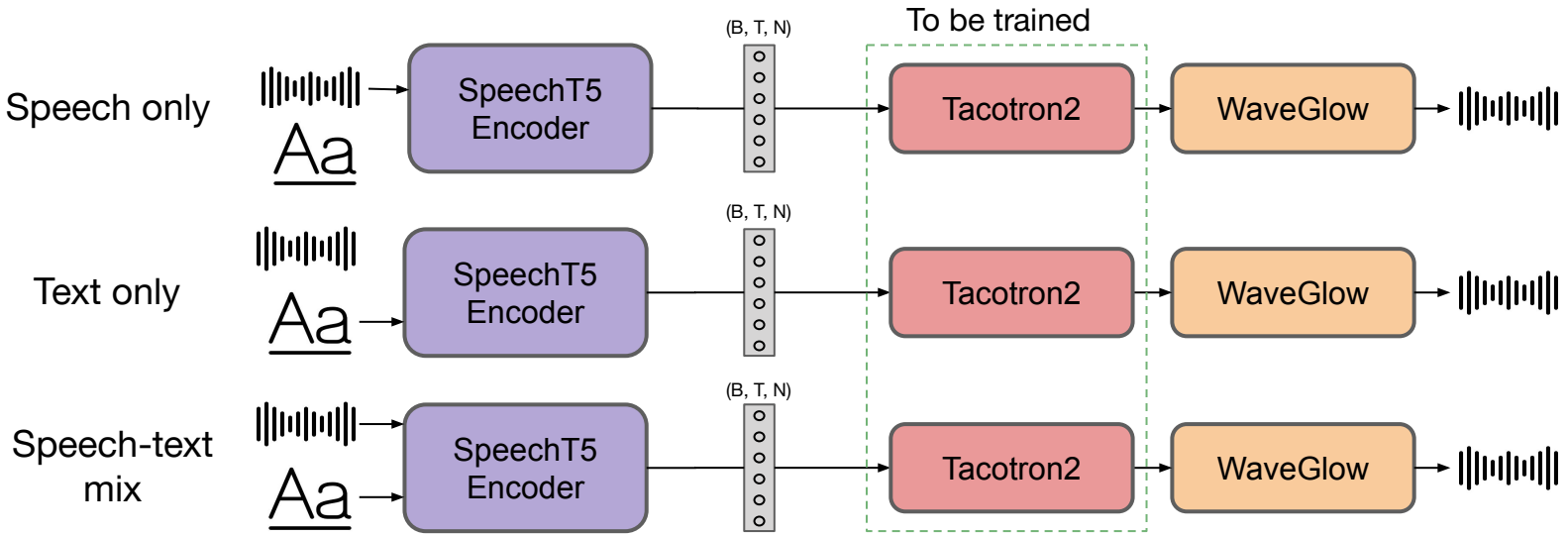
Decoding Team
SAMU-XLSR & LaBSE

- ➔ LaBSE hidden layer
  - Next step



Decoding Team
SpeechT5

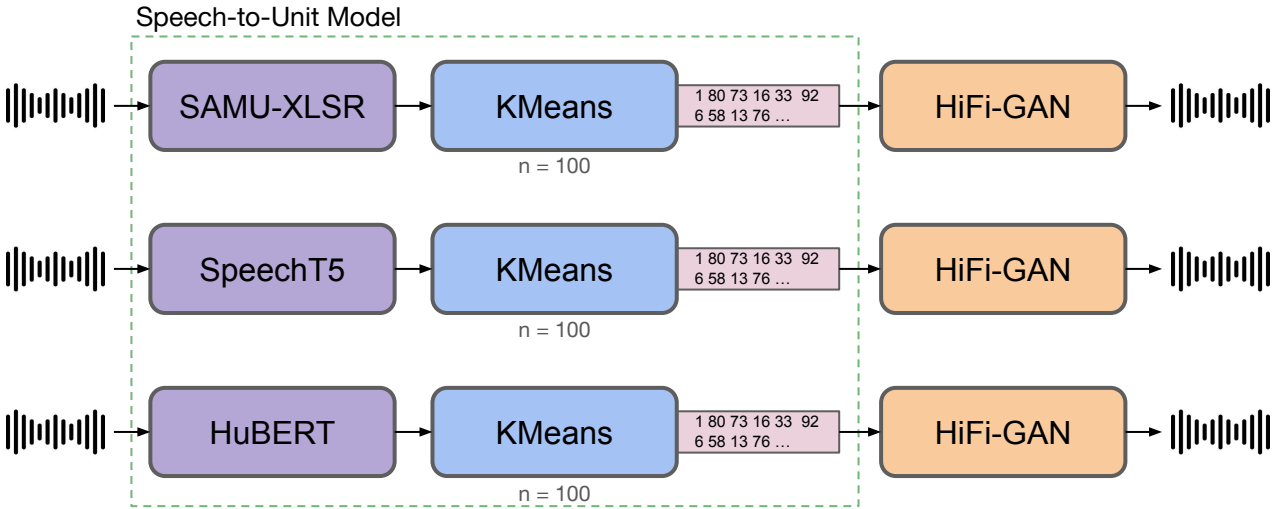
Objective : Compare synthetic speech from SAMU-XLSR/Labse (multilingual embeddings) and SpeechT5 (monolingual embeddings)



[1] J. Ao et al., " SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing," 2021. [Online]. Available: https://arxiv.org/abs/2110.07205

Decoding Team
Speech Generation

Objective : Generate speech directly from self-supervised discrete representations



[1] A. Polyak et al., " Speech Resynthesis from Discrete Disentangled Self-Supervised Representations" 2021. [Online]. Available: https://arxiv.org/abs/2104.00355

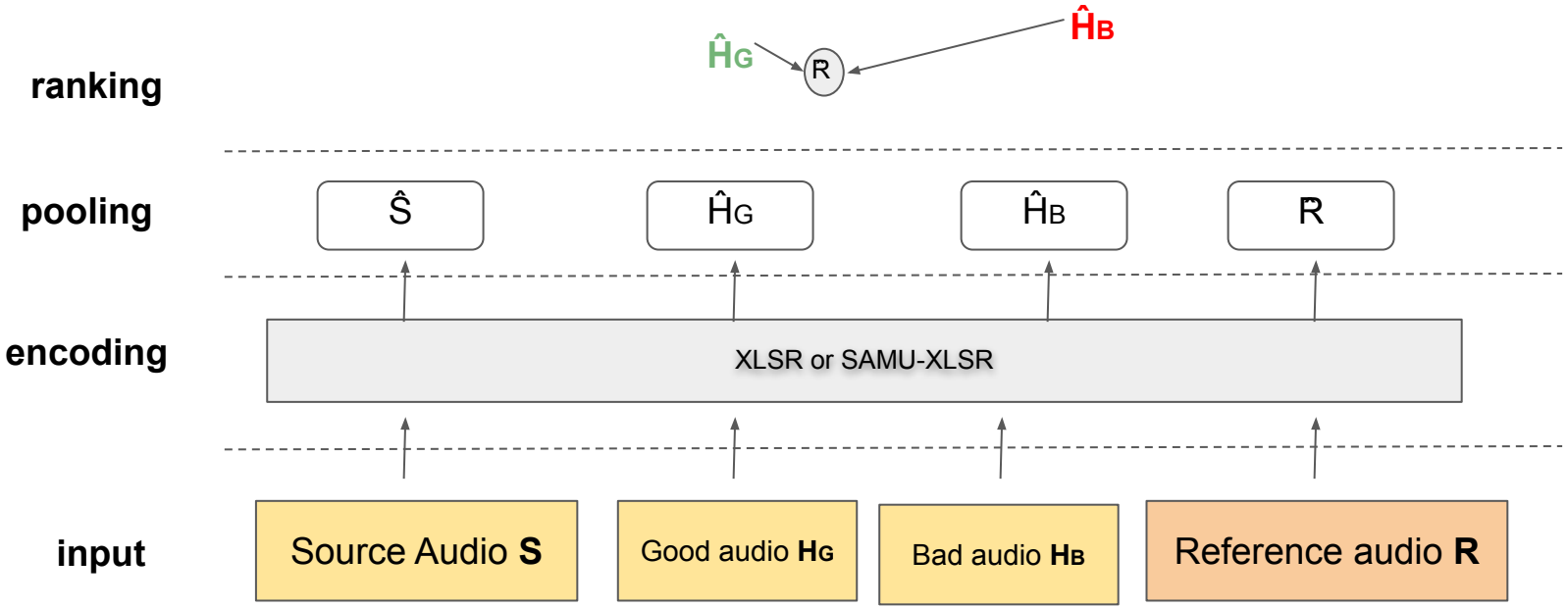
Decoding Team
Evaluation

How to evaluate generated speech ? *(possibly w/o availability of textual reference)*

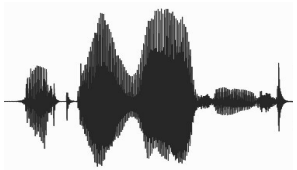
- A **textless** metric that compares a speech hypothesis (**H**) with a speech reference (**R**) along several axes
- The main axis is **meaning**, i.e similarity score should be high if both utterances convey the same message
- But other axes are interesting
  - eg. high similarity if H and R **voices** are similar (similar speaker, gender, etc.)

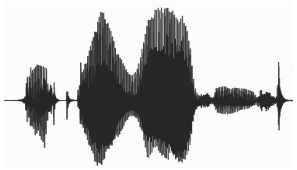
Decoding Team
Evaluation

A trainable metric, inspired by COMET (MT metric) architecture



Analysis Team





XLS-R

SAMU-XLSR

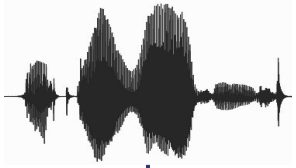
LaBSE

Direct Analysis

TEXT DECODER

TACOTRON2

DOWNSTREAM TASK



SLU / ASR / LID / ASV / ER / GID

Analysis Team

Insights about the first model produced :

- About the encoders : LaBSE / XLS-R / SAMU-XLSR
  - On downstream tasks : SLU / ASR / LID / ASV / ER
  - On the embeddings and the layers
- About the decoder : Tacotron / SpeechT5
  - On the reconstructed (translated) speech : ASV

Analysis Team
ASR and SLU analysis

Extraction of embeddings : Media (**FR**) and PortMedia (**IT**) using **SAMU-XLSR** and **XLS-R**

ASR and SLU on the embeddings to compute WER and CER (Concept Error Rate)

Evaluating the cross-language generalisation:

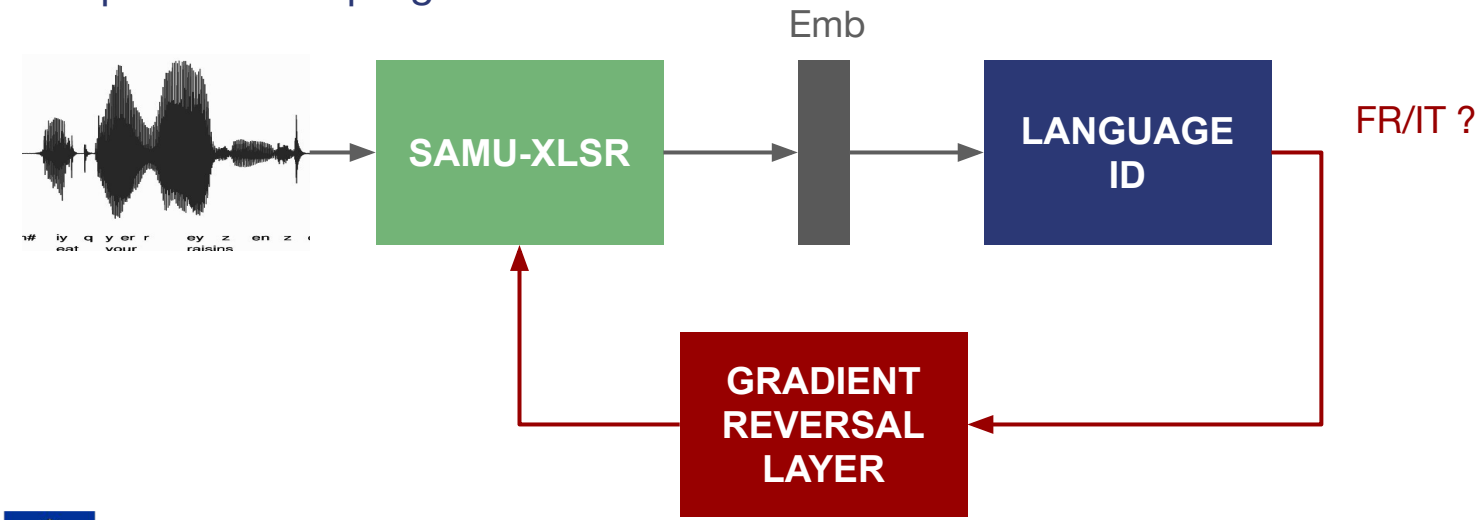
- Train FR → Test FR (baseline)
- Train FR → Test IT (zero-shot)
- Train FR + IT → Test IT (transfer learning)

Evaluating the cross-model performances

Analysis Team
Language ID : analysis and removal

Currently training : Gradient Reversal Layer to make FR and IT embeddings alike.

Experiments in progress :



Analysis Team
Language ID : analysis and removal

Can we use a Discriminator for an Italian embedding to be alike a normalisation of a typical French embedding to improve a latin module?

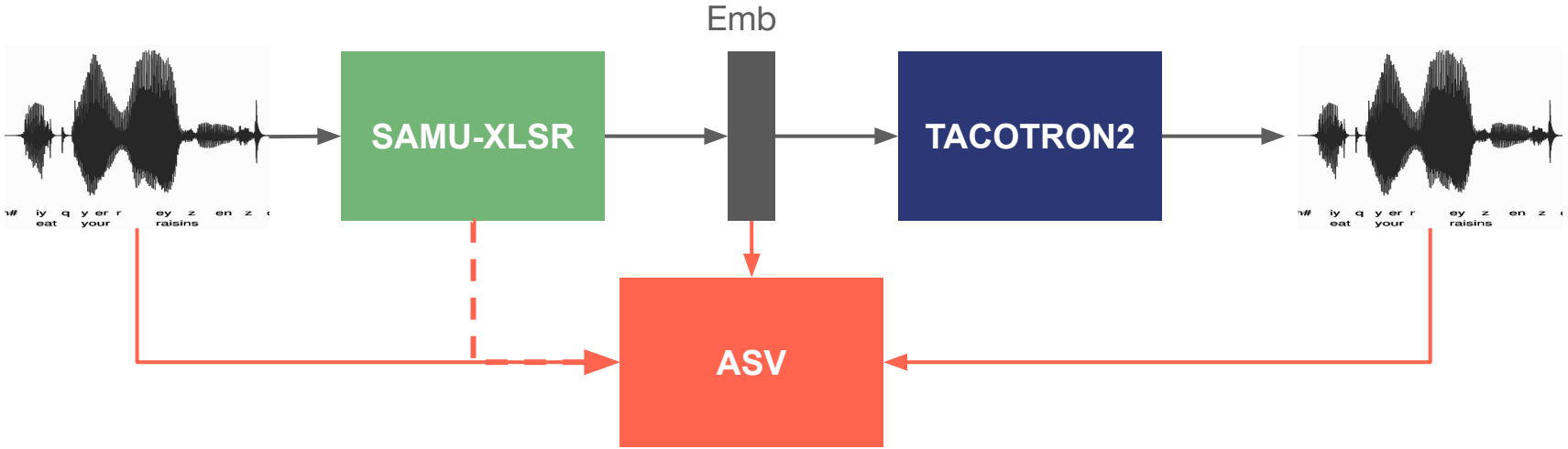
Some ideas…

- Gradient Reversal Layer with LID classification
  → might just reverse the labels and keeping LID information
- Normalize the IT embedding to be alike a standard FR embedding
  → will result in a confusion between FR and IT
- Adversarial Auto Encoder
- Reducing Domain Mismatch by Maximum Mean Discrepancy Based Autoencoders (*Weiwei LIN et al.)*

Analysis Team
ASV analysis

Extraction of embeddings : VoxCeleb1&2 and VCTK using SAMU-XLSR

Experiments in progress :



Analysis Team
ASV analysis

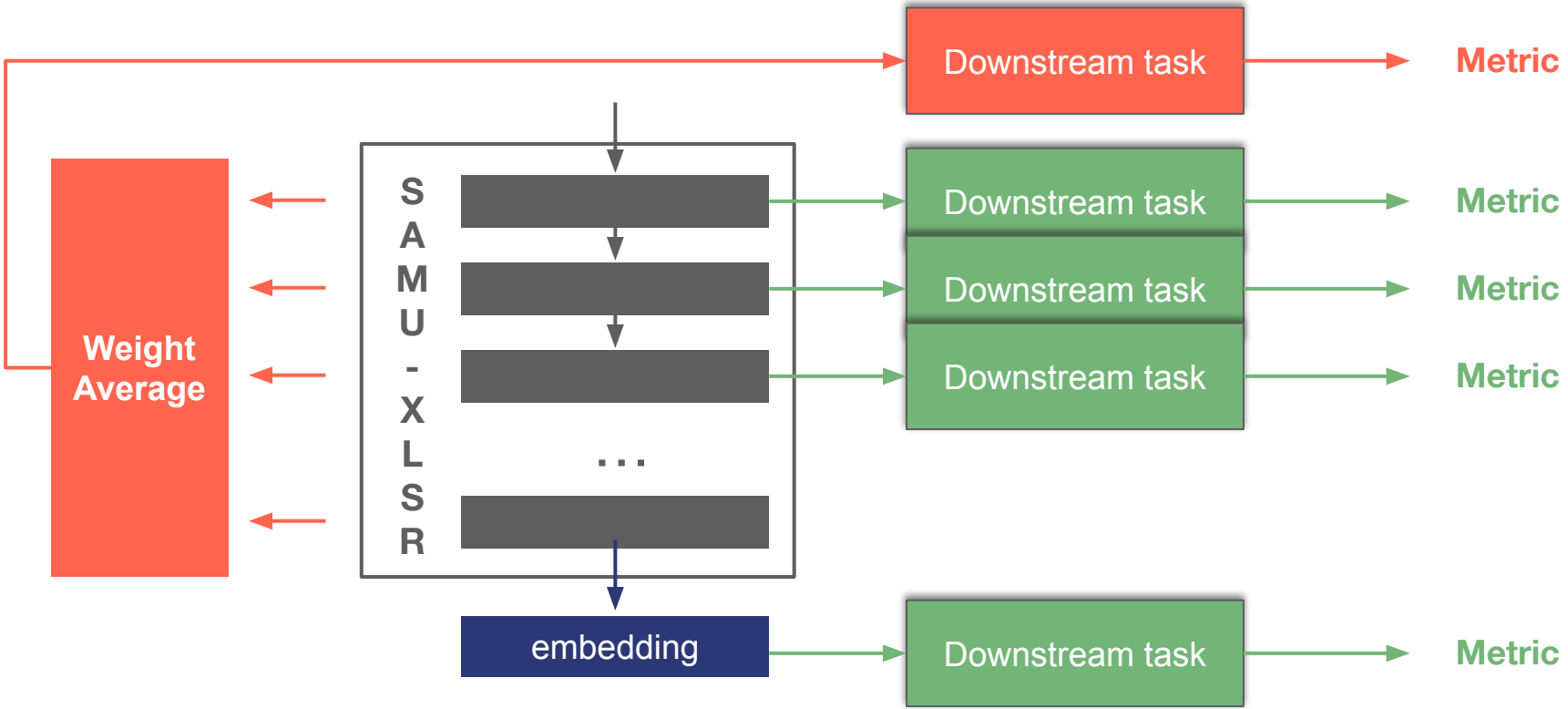
Extraction of embeddings : VoxCeleb1&2 and VCTK using SAMU-XLSR

Experiments in progress :

- Training ASV system
- Testing on source speech (baseline)
- Testing on intermediate representations (by SAMU-XLSR)
- Testing on reconstructed speech (by Various Encoders + Tacotron2)
- Testing on reconstructed (translated) speech (by Various Encoders + Tacotron2)

Analysis Team

Layer-wise probing task



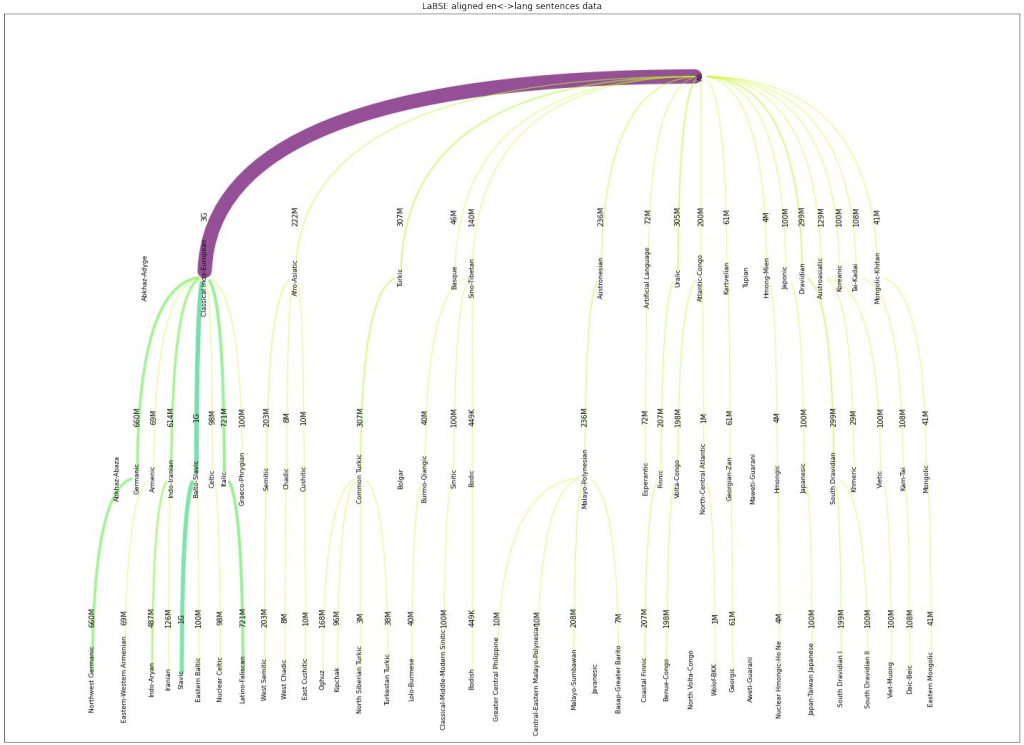
Esperanto
Exchanges for SPEech
ReseArch aNd TechnOlogies
Horizon 2020 project

Encoder+Analysis Team
Linguistic analysis

- Afro-Asiatic
  - Berber
    - Tuareg
      - Southern
        - **Tamashek**

- Indo-European
  - Italic
    - Latino-Faliscan
      - Romance
        - Western Romance
          - Gallo-Romance
            - Oïl
              - **French**

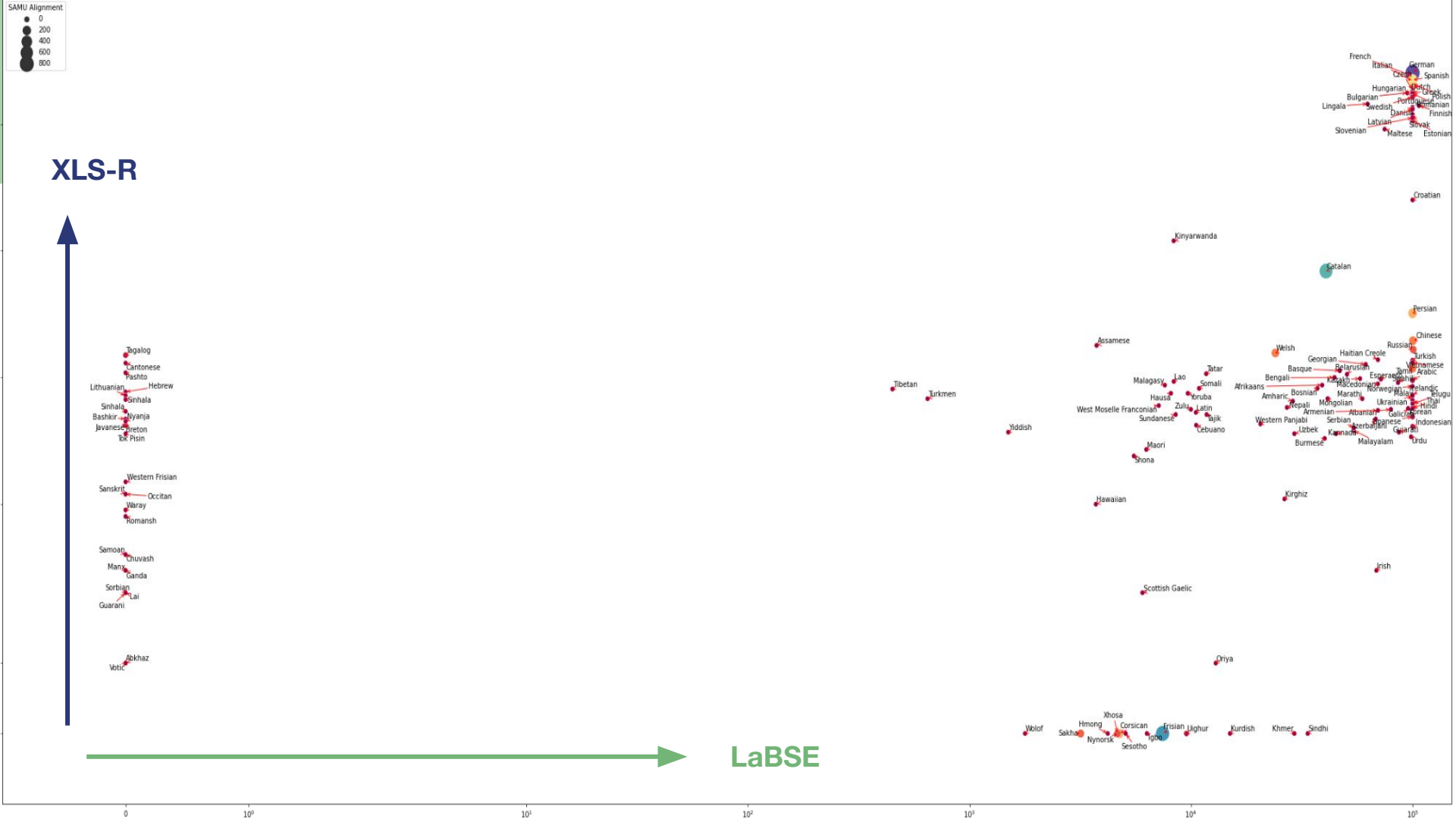
- Indo-European
  - Germanic
    - West Germanic
      - North Sea Germanic
        - Anglo-Frisian
          - Anglic
            - **English**



XLS-R :
128 Languages

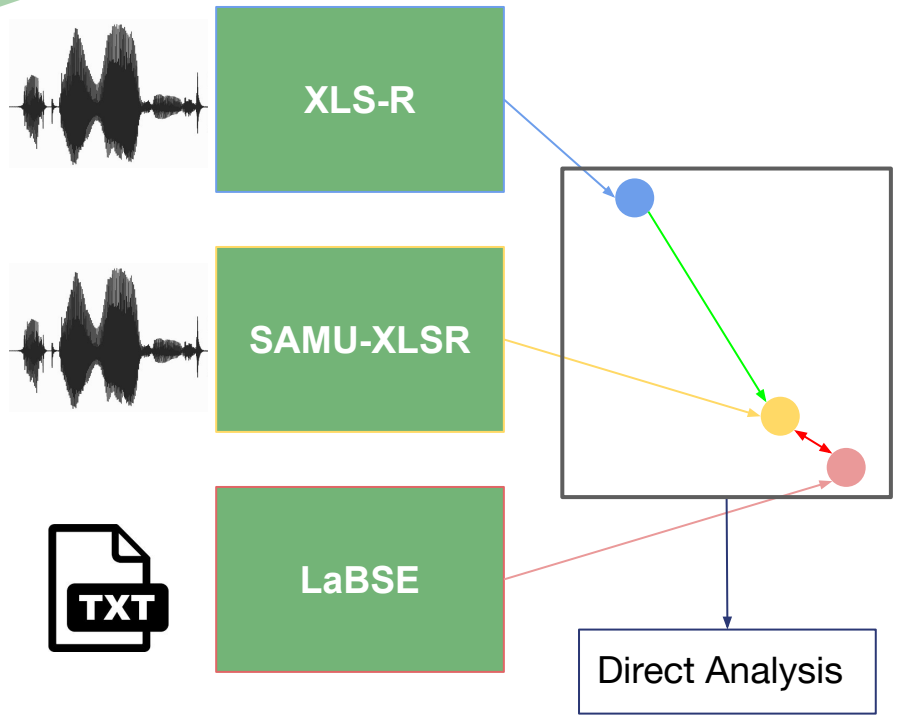
LaBSE :
101 Languages

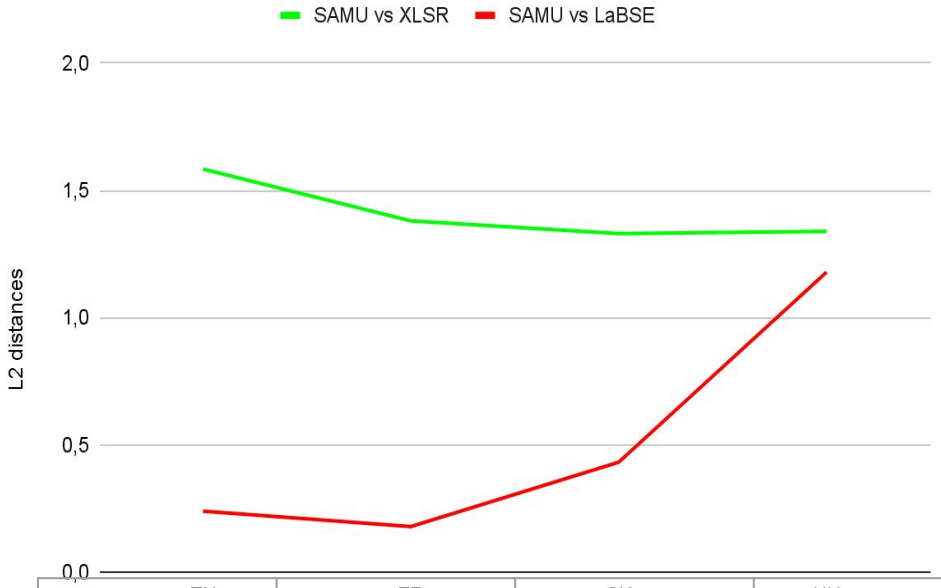
SAMU-XLSR :
51 Languages


SAMU Alignment
0
200
400
600
800
XLS-R
LaBSE
Tagalog
Cantonese
Pashto
Lithuanian
Hebrew
Sinhala
Sinhala
Bashkir
Nyanja
Javanese
Breton
Tok Pisin
Western Frisian
Sanskrit
Occitan
Waray
Romansh
Samoan
Chuvash
Manx
Ganda
Sorbian
Lai
Guarani
Abkhaz
Volapük
Tibetan
Turkmen
Yiddish
Assamese
Malagasy
Lao
Tatar
Somali
Hausa
Yoruba
Zulu
Latin
West Moselle Franconian
Sundanese
Tajik
Cebuano
Maori
Shona
Hawaiian
Kinyarwanda
Catalan
Persian
Croatian
French
Italian
German
Czech
Spanish
Hungarian
Dutch
Greek
Bulgarian
Portuguese
Polish
Lingala
Swedish
Romanian
Danish
Finnish
Latvian
Slovak
Slovenian
Maltese
Estonian
Chinese
Welsh
Russian
Haitian Creole
Turkish
Georgian
Vietnamese
Basque
Belarusian
Arabic
Tamil
Bengali
Esperanto
Kazakh
Afrikaans
Macedonian
Icelandic
Amharic
Bosnian
Norwegian
Telugu
Nepali
Marathi
Malay
Mongolian
Ukrainian
Thai
Armenian
Albanian
Gaelic
Hindi
Western Panjabi
Serbian
Japanese
Korean
Uzbek
Azerbaijani
Indonesian
Kannada
Gujarati
Burmese
Malayalam
Urdu
Kirghiz
Irish
Scottish Gaelic
Oriya
Xhosa
Wolof
Hmong
Corsican
Frisian
Uighur
Kurdish
Khmer
Sindhi
Sakha
Nynorsk
Igbo
Sesotho
0
10^0
10^1
10^2
10^3
10^4
10^5

Esperanto
Exchanges for SPEech ReseArch aNd TechnOlogies
Horizon 2020 project

Analysis Team
Embeddings analysis





EN (EN)	FR (FR)	SK (SK)	HU (HU)
Indo european			Uralic
Germanic	Italic	Balto-slav	Finno-Ugric
Used for SAMU		Not Used for SAMU	

Questions?

Esperanto

Exchanges for SPEech
ReseArch aNd TechnOlogies

Horizon 2020 project