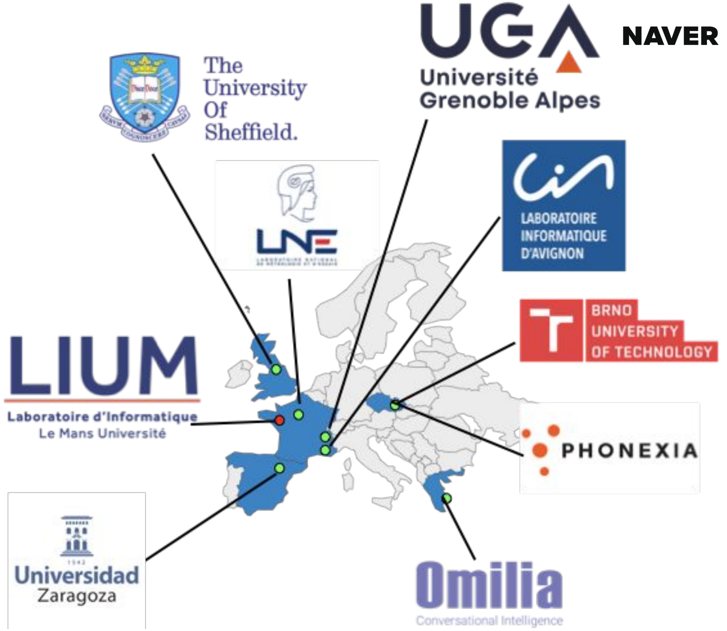
Multi-lingual Speech to Speech Translation for Under-Resourced languages

Esperanto

Exchanges for SPEech ResearchAND TechnOlogies

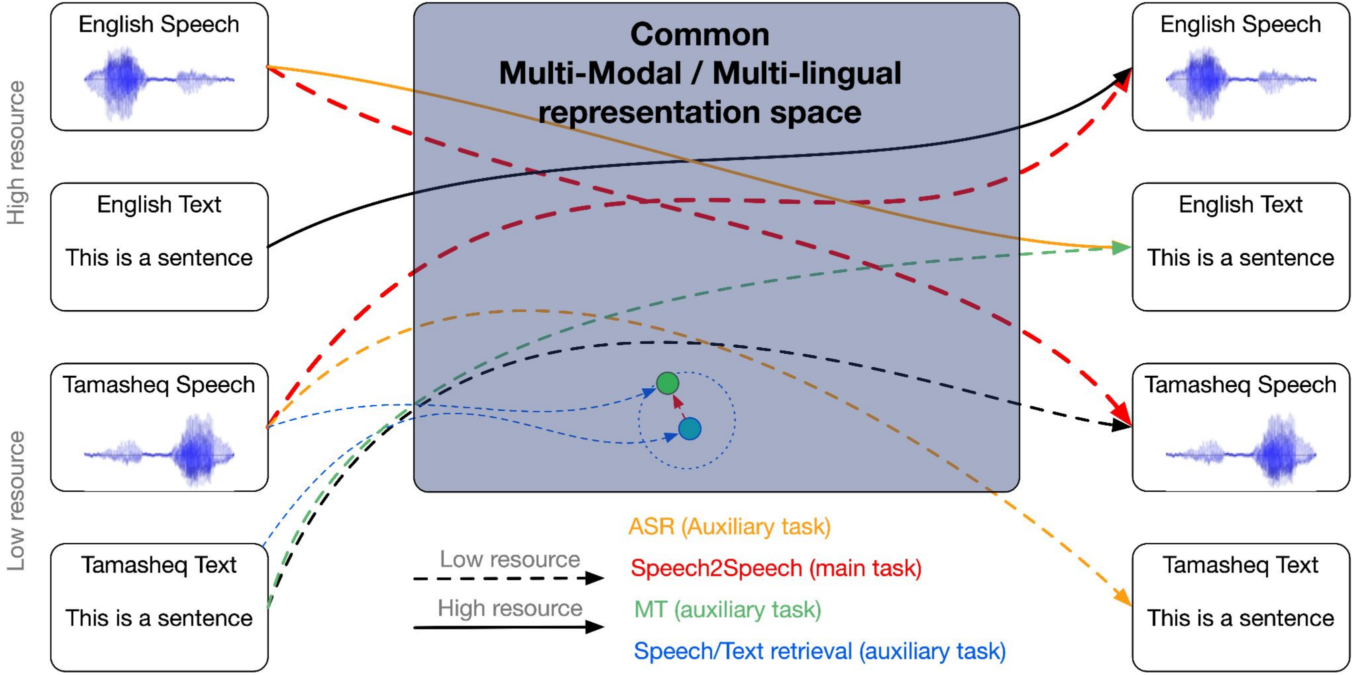
Horizon 2020 project

The TEAM


UGA
Université Grenoble Alpes
NAVER LABS
Europe
The University Of Sheffield.
LNE
LABORATOIRE INFORMATIQUE D'AVIGNON
BRNO UNIVERSITY OF TECHNOLOGY
LIUM
Laboratoire d'Informatique
Le Mans Université
PHONEXIA
Universidad Zaragoza
Omilia
Conversational Intelligence

Esperanto
Exchanges for SPEech
ReseArch aNd TechnOlogies
Horizon 2020 project

The Goal



The Goal

Develop a Multi-Modal / Multi-Lingual / Extensible Translation system

Multi-Modal

- Text / Speech inputs
- Text / Speech outputs

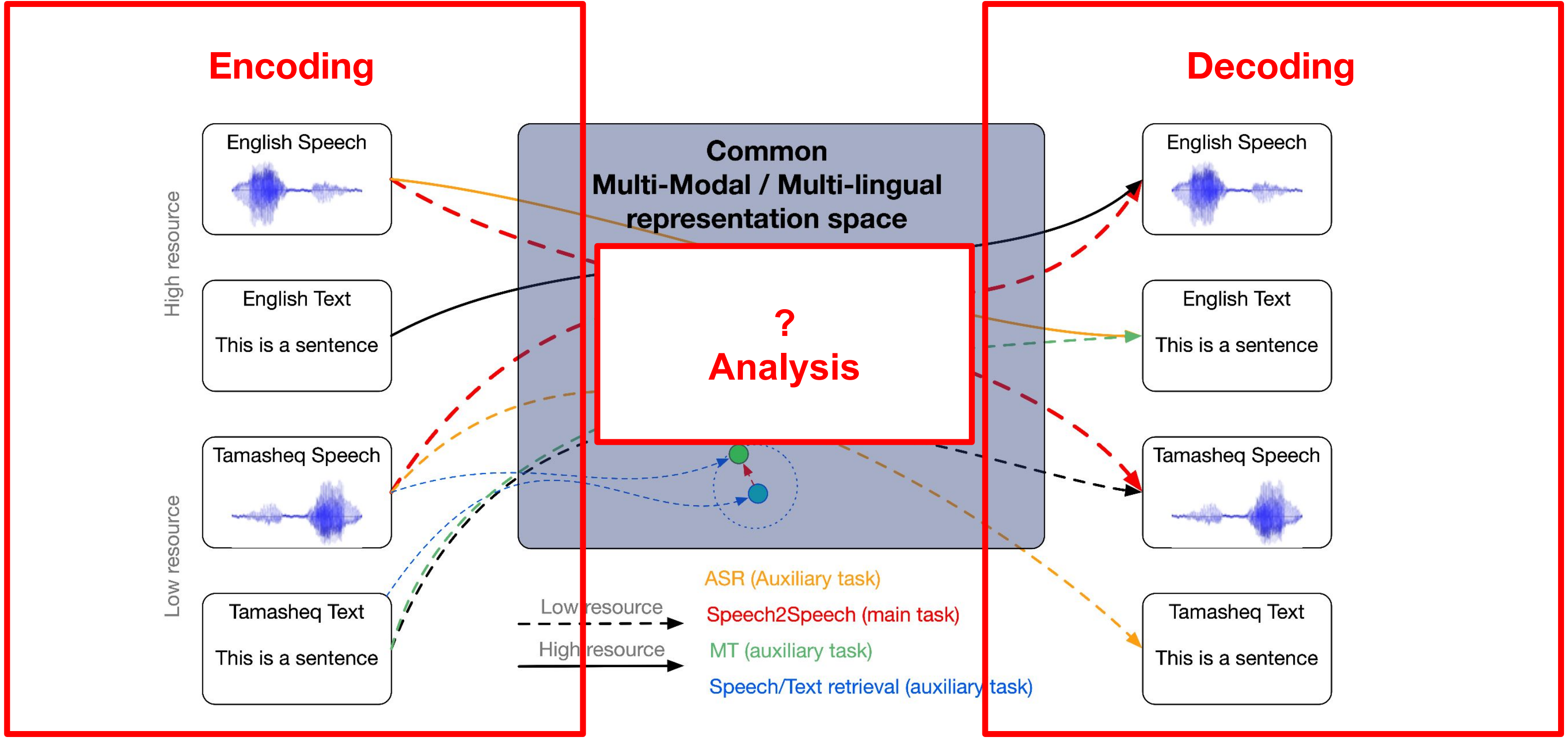
Multi-Lingual

- Assume the existence of a common multi-lingual space

Extensible

- Easily add new languages with low resources
- Voice conversion / anonymisation / pseudo-anonymisation

The Organization



Encoding Team
Goals

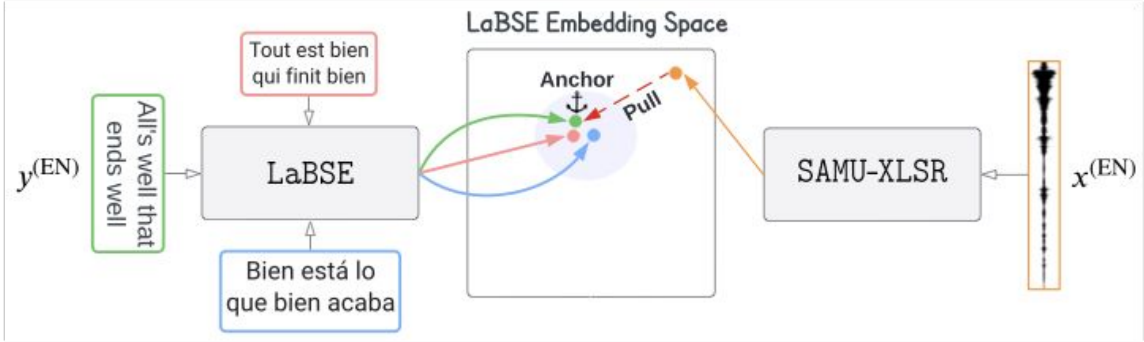
- Learn a multilingual semantically aligned semantic space (like labSE [1] and LASER [2] but for speech)
- XLSR does not project semantically aligned sentences in the same space
- This kind of encoder should transfer better to unseen languages for speech translation
- New architectures for pre-training multilingual LMs.

[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

Encoding Team
Challenges

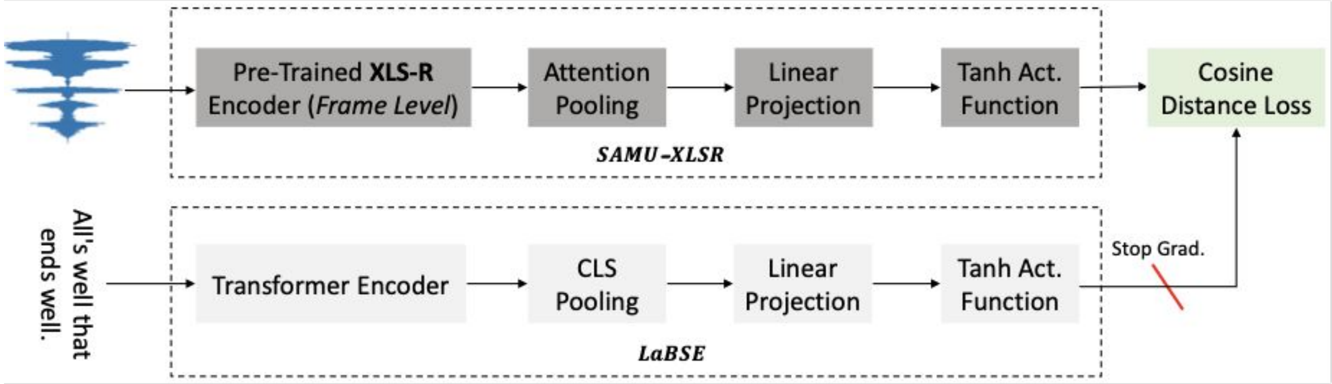
- We have a system working for speech retrieval (SAMU-XLSR [3])


LaBSE Embedding Space
Tout est bien qui finit bien
All's well that ends well
$y^{(EN)}$
LaBSE
Bien está lo que bien acaba
Anchor
Pull
SAMU-XLSR
$x^{(EN)}$

[3] S. Khurana, A. Laurent, J. Glass, SAMU-XLSR: Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation, IEEE Journal of Selected Topics in Signal Processing

Encoding Team
Challenges

- How to make it output a sequence of embeddings that the decoder can use?



[3] S. Khurana, A. Laurent, J. Glass, SAMU-XLSR: Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation, IEEE Journal of Selected Topics in Signal Processing

Encoding Team
Challenges

- Fusion of monolingual (or language family based-) wav2vec2.0 models to address a new low-resourced language
  - Assumption 1: speech representations trained on a huge amount of languages lose precision
  - Assumption 2: multilingual SSL models are not suited to handle phonotactics that is mainly language-dependent

Decoding Team
Goals

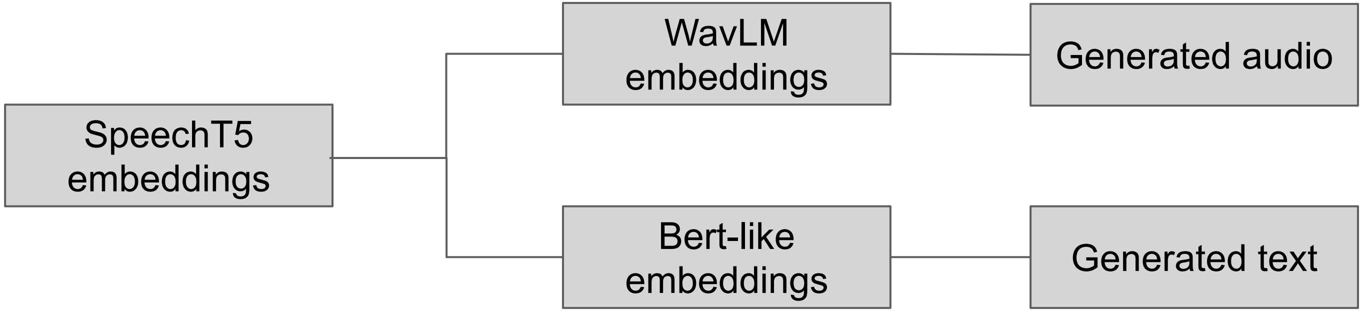
- Generate text and speech from the encoded data
- Common representation as an input > Need to divide the information into speech- and text-related parts
- Depends on what information remains in the encoded space
- Evaluate audio-only outputs (speech2speech metrics)

Decoding Team
Challenges

- How to divide speech and text information ?
- Can we jointly decode speech and text ?
- How can we choose the target language ?
- Can we control speaker information while decoding audio output ?
- How can we evaluate generated speech and text ?

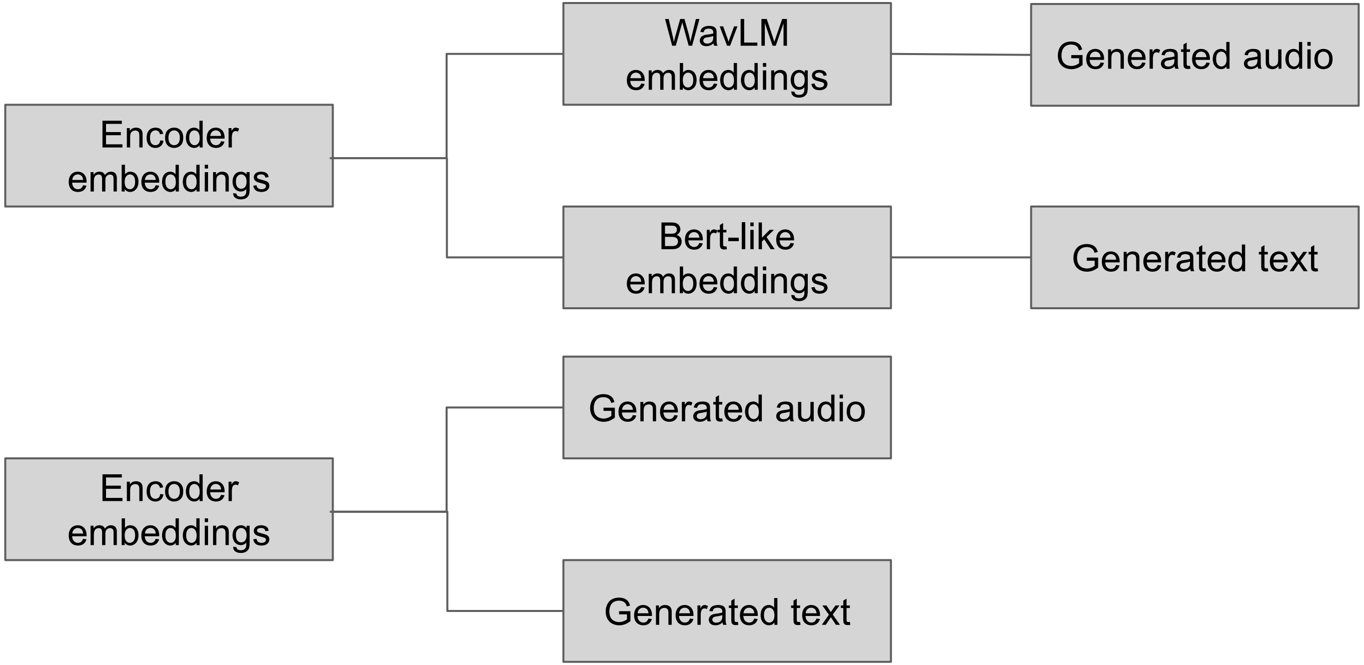
Decoding Team

One possible starting point: generate speech and text representation sequences from multimodal embeddings



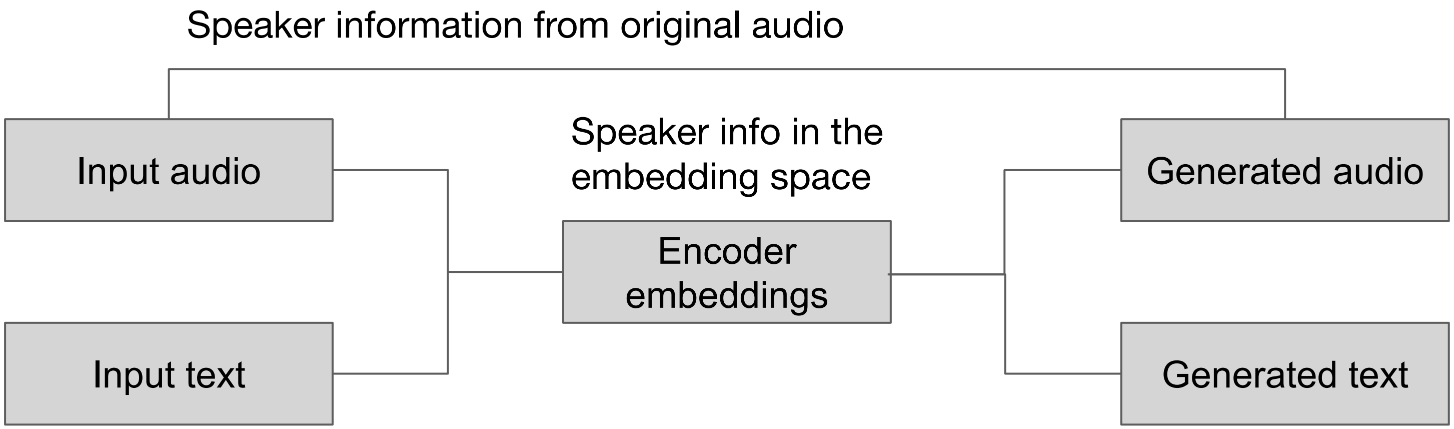
Decoding Team

Then, depending on the results of the Encoder team and of our previous experiments, see how we can merge both systems



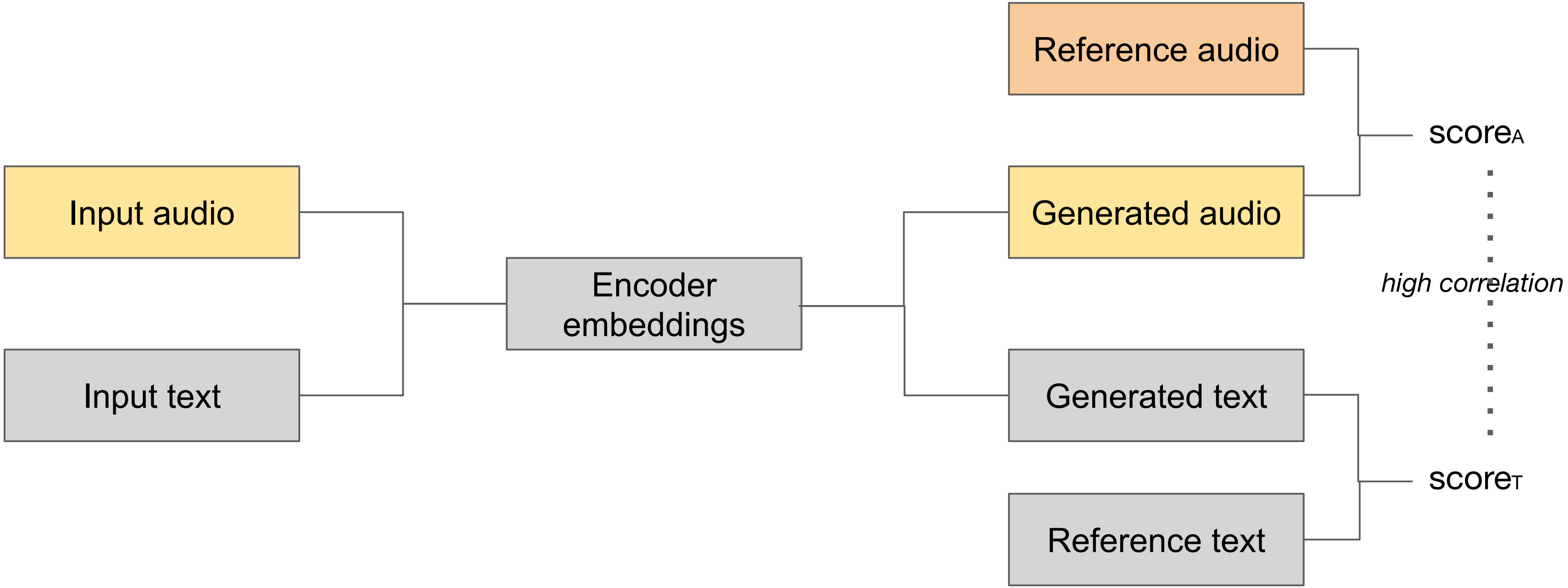
Decoding Team

How to control some aspects of generated voice ?

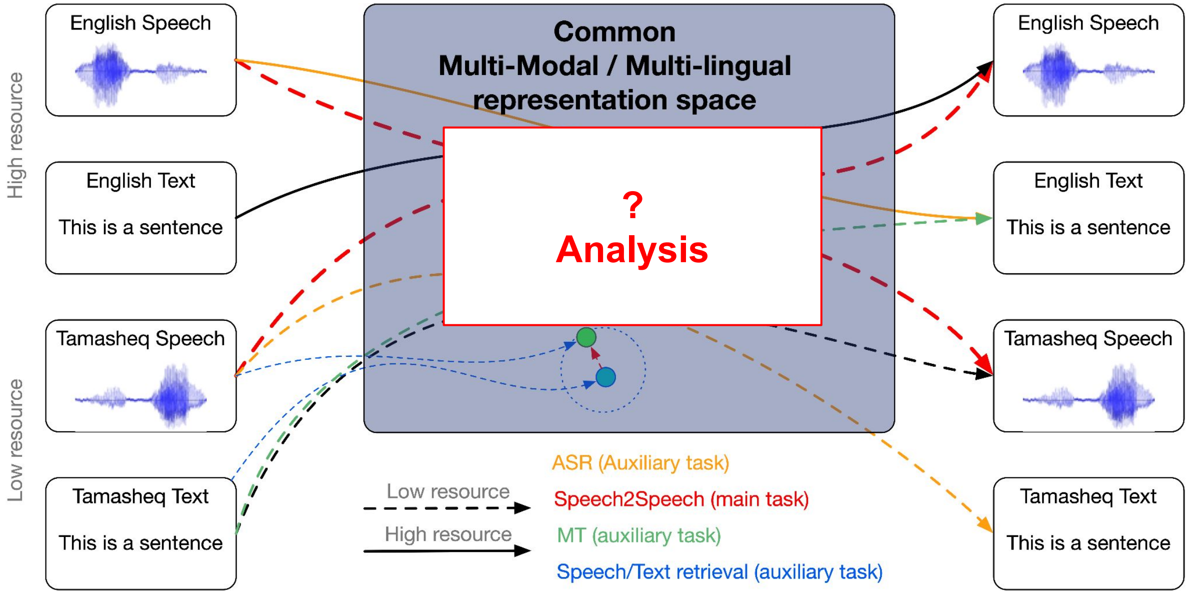

Speaker information from original audio
Input audio
Input text
Speaker info in the embedding space
Encoder embeddings
Generated audio
Generated text

Decoding Team

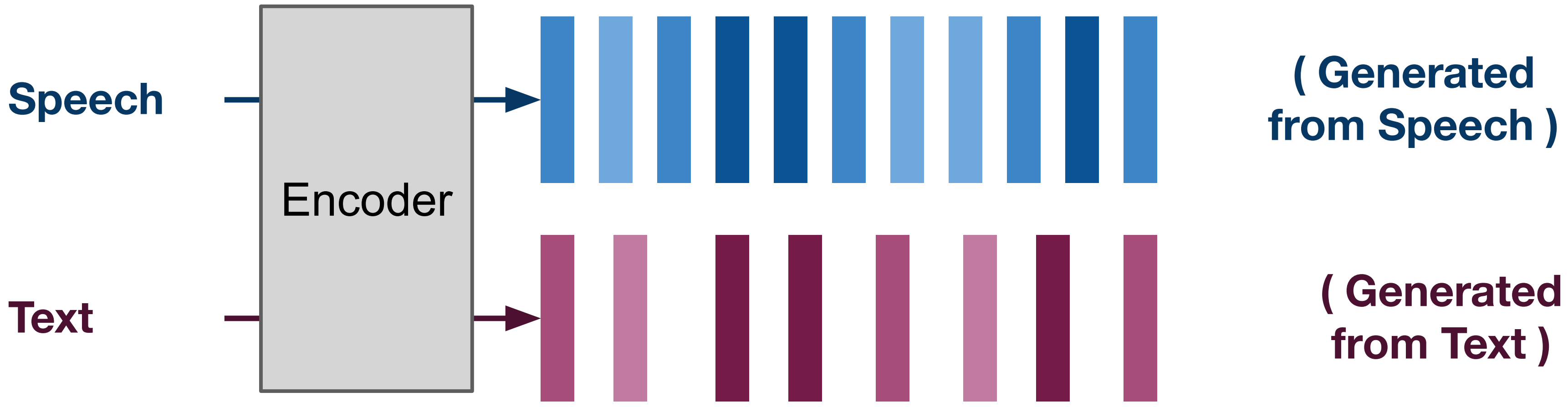
How to evaluate generated speech ? *(possibly w/o availability of textual reference)*



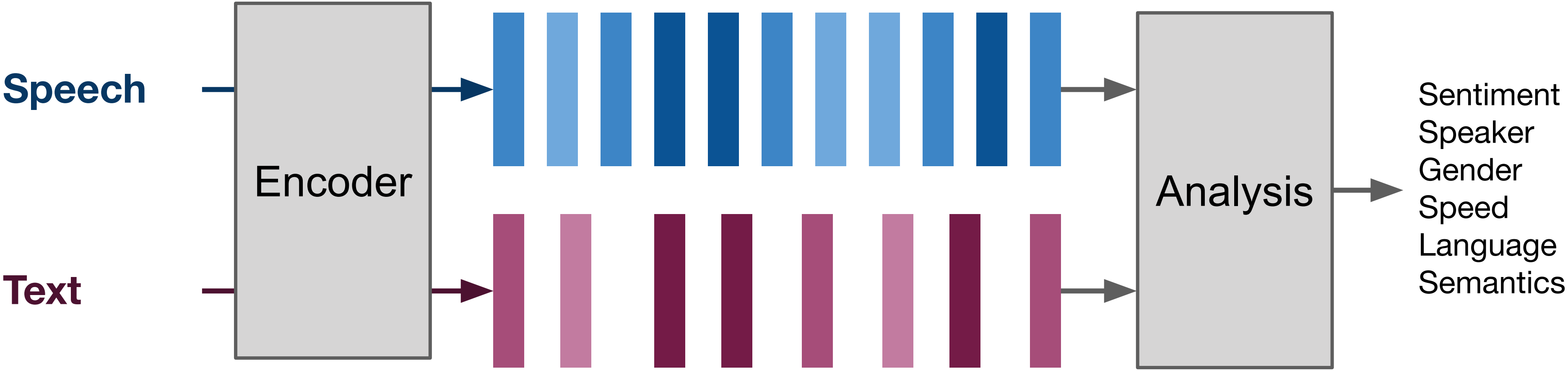
Analysis Team



Analysis Team

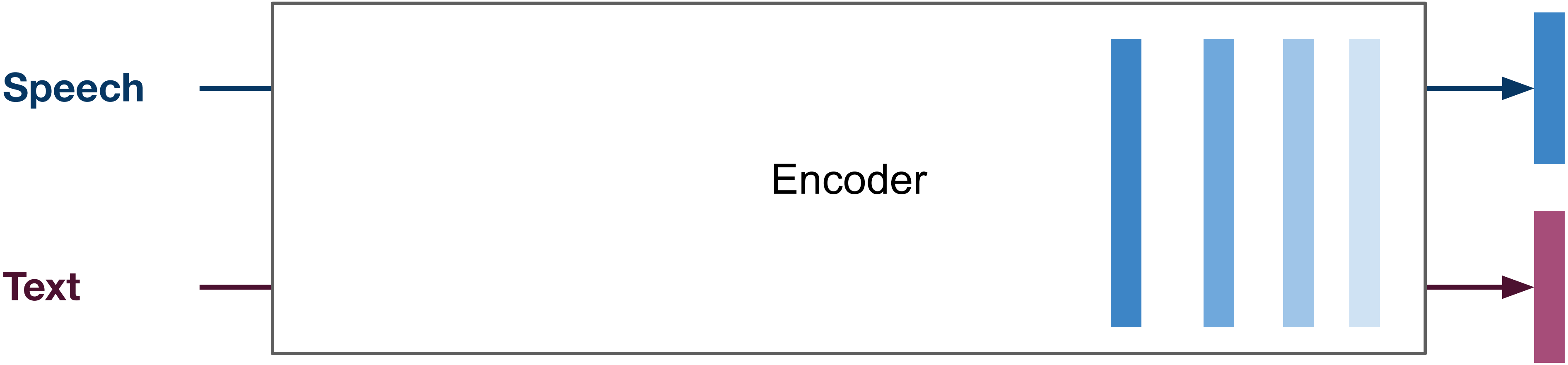


Analysis Team

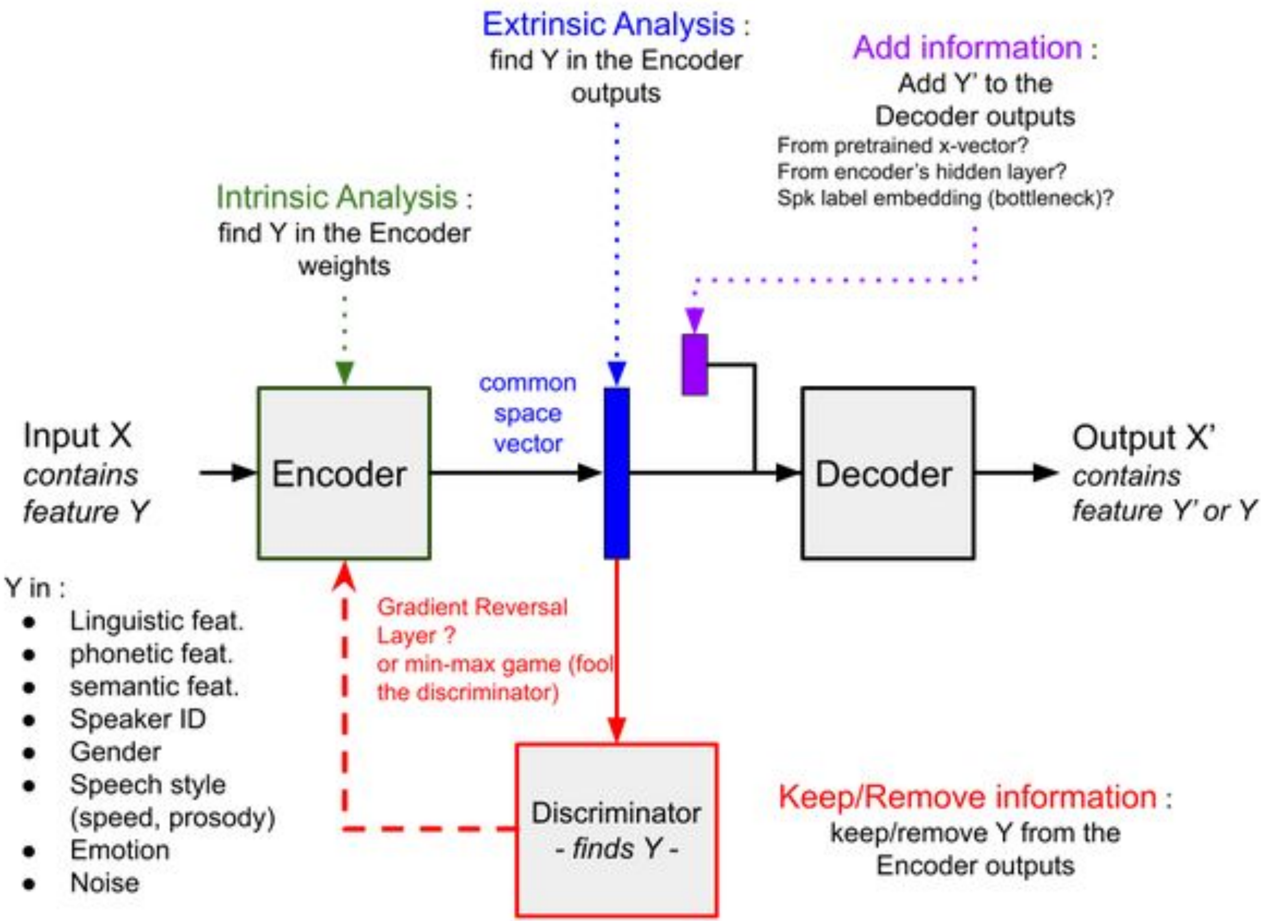


Extrinsic Probing sub-task

Analysis Team



Intrinsic Probing sub-task



Extrinsic Analysis : find Y in the Encoder outputs

Add information : Add Y' to the Decoder outputs

From pretrained x-vector?
From encoder's hidden layer?
Spk label embedding (bottleneck)?

Intrinsic Analysis : find Y in the Encoder weights

Input X *contains feature Y*

Encoder

common space vector

Decoder

Output X' *contains feature Y' or Y*

Y in :

- Linguistic feat.
- phonetic feat.
- semantic feat.
- Speaker ID
- Gender
- Speech style (speed, prosody)
- Emotion
- Noise

Gradient Reversal Layer ?
or min-max game (fool the discriminator)

Discriminator
- finds Y -

Keep/Remove information : keep/remove Y from the Encoder outputs

Analysis Team

The Data

- VoxCeleb (EN Speakers)
- MEDIA/PortMEDIA (FR/IT Semantics)
- VoxPopuli/CommonVoice (Multi Languages Speech)
- Librispeech/MULTIATIS++ (Multimodal EN)
- MELD/IEMOCAP (Emotions)

Analysis Team

Done List

- List of targets
- Attentive pooling for improved ER / ASV
- Removing language ID with gradient reversal (for ASR/SLU tasks)
- Statement : We need a day-to-day follow up on others tasks advances

Analysis Team

TODO List

- □ Agree on common baselines
- □ Setup the gitlab/framework/architecture
- □ Mesure performances for the baseline systems
- □ Make all the metrics automatic
- □ Influence the network to add/remove precise information
- □ Analyze the embeddings generated by “all” layers in different ways

Multi-lingual Speech to Speech Translation for Under-Resourced languages

Esperanto

Exchanges for SPEech
ResеArch aNd TechnOlogies
Horizon 2020 project