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

Esperanto
Exchanges for SPEech Research and TechnOlogies
Horizon 2020 project
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101007666 / ESPERANTO / H2020-MSCA-RISE-2020.

The TEAM
The Goal

Common Multi-Modal / Multi-lingual representation space

- English Speech
- English Text
- This is a sentence
- Tamasheq Speech
- Tamasheq Text
- This is a sentence

ASR (Auxiliary task)
Speech2Speech (main task)
MT (auxiliary task)
Speech/Text retrieval (auxiliary task)
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
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101007666 / ESPERANTO / H2020-MSCA-RISE-2020.

The Organization

Encoding

- English Speech
- English Text
  - This is a sentence
- Tamasheq Speech
- Tamasheq Text
  - This is a sentence

Decoding

- English Speech
- English Text
  - This is a sentence
- Tamasheq Speech
- Tamasheq Text
  - This is a sentence

Common Multi-Modal / Multi-lingual representation space

Analysis

ASR (Auxiliary task)
Speech2Speech (main task)
MT (auxiliary task)
Speech/Text retrieval (auxiliary task)
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.


Encoding Team

Challenges

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

Encoding Team Challenges

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

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

- SpeechT5 embeddings
- WavLM embeddings
- Bert-like embeddings
- Generated audio
- Generated text
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101007666 / ESPERANTO / H2020-MSCA-RISE-2020.

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

Decoding Team

WavLM embeddings → Generated audio

Bert-like embeddings → Generated text

Generated audio → Generated text

Generated audio → Generated text

Encoder embeddings → WavLM embeddings

Encoder embeddings → Bert-like embeddings

Encoder embeddings → Generated audio

Encoder embeddings → Generated text
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)*

- Input audio
- Input text
- Encoder embeddings
- Reference audio
- Generated audio
- Generated text
- Reference text

Score $A$, $T$ high correlation
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101007666 / ESPERANTO / H2020-MSCA-RISE-2020.

Analysis Team

Common Multi-Modal / Multi-lingual representation space

? Analysis

- English Speech
- English Text
  - This is a sentence
- Tamasheq Speech
- Tamasheq Text
  - This is a sentence

- ASR (Auxiliary task)
- Speech2Speech (main task)
- MT (auxiliary task)
- Speech/Text retrieval (auxiliary task)
Analysis Team

Speech

Encoder

Text

(Generated from Speech)

(Generated from Text)
Analysis Team

Extrinsic Probing sub-task
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101007666 / ESPERANTO / H2020-MSCA-RISE-2020.

Analysis Team

Speech

Text

Intrinsic Probing sub-task
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101007666 / ESPERANTO / H2020-MSCA-RISE-2020.
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
- Measure 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