IWSLT 2022

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

ON-TRAC Consortium Systems for the IWSLT 2022 Dialect and Low-resource Speech Translation Tasks

Marcely Zanon Boito[1], John Ortega[2], Hugo Riguidel[2], Antoine Laurent[2], Loïc Barrault[2], Fethi Bougares[3], Firas Chaabani[3], Ha Nguyen[1,5], Florentin Barbier[4], Souhir Gahbiche[4], Yannick Estève[1]

[1]LIA - Avignon University, [2]LIUM - Le Mans University, [3]ELYADATA, [4]Airbus, [5]LIG - Grenoble Alpes University

DIALECT SPEECH TRANSLATION

Dataset:

- **383h** of transcribed **Tunisian Arabic** speech from which, **160h** are translated into **English**. The speech style is conversational speech;
- **1,200h** of **Modern Standard Arabic** (MSA) transcriptions (MGB2).

Primary System:

- Joint submission to low-resource and dialect tracks;
- E2E 6-layers encoder-decoder conformer model;
- Speed perturbation and SpecAugment were used. The model is trained on mel filterbanks (80 channels).

Contrastive Systems:

- The contrastive systems are cascade models.
- They both use a XLSR-53-based E2E ASR model trained on MSA and fine-tuned on Tunisian Arabic;
- Submitted models have no LM rescoring (45.1 WER), post-evaluation uses a 5-gram LM (41.5 WER);
- Two MT models are considered: a 4-layer bi-LSTM model (contrastive 1), and a 15-layer fully convolutional architecture (contrastive 2).

Results:

System	Description	valid	test
primary	E2E	12.2	12.4
contrastive 1	cascade	15.1	**13.6**
contrastive 2	cascade	12.8	11.3
post-evaluation	cascade	16.0	**14.4**

Table: BLEU4 results for the submitted systems (purple), and the post-evaluation complementary model (green).

Main Takeaways:

1. **Noise is also a challenge.** While this track focused on the dialect aspect, we find that the main challenge for our ASR models is producing transcriptions in noisy settings. The dataset is unfortunately of mixed quality in terms of background noise.

2. **E2E vs Cascade.** Our best model is the post-evaluation model utilizing a XLSR-53-based ASR model. We believe we could probably push E2E results further by integrating pre-trained Self-Supervised Learning models (e.g. wav2vec 2.0, HuBERT).

Information about the paper and References:

LOW-RESOURCE SPEECH TRANSLATION

Dataset:

- **17h** of **Tamasheq** audio translated into **French** text. The speech style is radio broadcast;
- **224h** of Tamasheq audio (with no annotation);
- **417h** in *geographically close languages* (French from Niger, Fulfulde, Hausa and Zarma).

Primary System:

- Pre-trained wav2vec 2.0 base model trained on Tamasheq only (241h),
- E2E ST model comprising a wav2vec 2.0 as encoder, a projection layer (768>256), and a Transformer decoder (3 layers, 4 heads, 256 dim).
- Best results achieved by removing the last 6 layers inside the wav2vec 2.0 encoder (W2V-6).

Contrastive System:

- Pre-trained French phonemic ASR model for generating *approximate transcriptions* in Tamasheq.
- E2E ST model comprising a conformer encoder (12 layers, 1024 dim), and a transformer decoder (3 layers, 2048 dim).
- The model is **jointly optimized** on ST, ASR and MT tasks.

Results:

System	Description	valid	test
primary	E2E, W2V-6+ST	8.3	**5.7**
contrastive	E2E, ASR+ST	6.4	5.0
contrastive 2	pipeline, W2V-ASR+ST	3.6	3.1
contrastive 3	pipeline, W2V-FT+ST	2.9	2.5
baseline	pipeline	2.2	1.8

Table: BLEU4 results for the submitted systems (purple), and post-evaluation complementary models (green).

Main Takeaways:

1. **Intermediate representations** from wav2vec 2.0 are useful for low-resource settings:
   - Higher abstraction level for the speech,
   - Less parameters to train.

2. **Approximate transcriptions** produced by ASR models in different languages can be used for constraining ST models.

3. **Poor results for off-the-shelf wav2vec 2.0 models,** even after fine-tuning on target data. Also, we do not observe good results using wav2vec 2.0 as feature extractor (contrastive 2 and contrastive 3).