ON-TRAC' systems for the IWSLT 2021 low-resource speech translation and multilingual speech translation shared tasks

Hang Le[2], Florentin Barbier[4], Ha Nguyen[1,2], Natalia Tomanshenko[1], Salima Mdhaffar[1], Souhir Gahbiche[4], Fethi Bougares[3], Benjamin Lecouteux[2], Didier Schwab[2], Yannick Estève[1]
ON-TRAC consortium (LIA[1], LIG[2], LIUM[3], Airbus[4] - France)

Abstract

This paper describes the ON-TRAC Consortium translation systems developed for two challenge tracks featured in the Evaluation Campaign of IWSLT 2021, low-resource speech translation and multilingual speech translation. The ON-TRAC Consortium is composed of researchers from three French academic laboratories and an industrial partner: LIA (Avignon Université), LIG (Université Grenoble Alpes), LIUM (Le Mans Université), and researchers from Airbus. A pipeline approach was explored for the low-resource speech translation task, using a hybrid HMM/TDNN automatic speech recognition system fed by wav2vec features, coupled to an NMT system. For the multilingual speech translation task, we investigated the use of a dual-decoder Transformer that jointly transcribes and translates an input speech signal. This model was trained in order to translate from multiple source languages to multiple target ones.

1 Introduction

In the two last editions of the IWSLT evaluation campaigns, the ON-TRAC consortium focused on end-to-end offline speech translation and simultaneous speech translation (Nguyen et al., 2019; Elbayad et al., 2020). In 2021, we chose to focus on low-resource speech translation and multilingual speech translation by using two different kinds of approaches: a cascaded speech-to-text translation (combining source language automatic speech recognition (ASR) and source-to-target text translation) to process the low resource speech translation tasks, and a neural end-to-end model for the multilingual speech translation task. For the low resource task, we investigated the use of speech features extracted by a neural model pretrained by self supervision the wav2vec XLSR-53 model (Conneau et al., 2020) in order to process Swahili languages by a classical hybrid Markovian/neuronal ASR system. The ASR outputs were processed by neural machine text-to-text translation systems dedicated to the two targeted language pairs. For the multilingual speech translation task, we investigated the use of a dual-decoder Transformer that jointly transcribes and translates an input speech. This model was trained in order to translate from multiple source languages to multiple target ones.

The ON-TRAC Consortium is composed of researchers from three French academic laboratories and an industrial partner: LIA (Avignon Université), LIG (Université Grenoble Alpes), LIUM (Le Mans Université), and researchers from Airbus.

2 Low resource speech translation

The task of the low resource speech translation track was to build the speech transcription/translation system for transcribing and/or translating between the two language pairs:

- Coastal Swahili (swa) to English (eng)
- Congolese Swahili (swc) to French (fra)

2.1 ASR system

The same ASR models were used for both test datasets: Coastal Swahili (swa) and Congolese Swahili (swc).

2.1.1 Data

The training corpus for the ASR acoustic model (AM) comprises of several datasets:

- 5k instances for Congolese Swahili speech provided by the IWSLT-2021 organizers[1];
- a training subset of the ALFFA corpus (Gelas et al., 2012) (read speech and broadcast news);

[1] https://iwslt.org/2021/low-resource

Proceedings of the 18th International Conference on Spoken Language Translation, pages 169–174
Bangkok, Thailand (Online), August 5–6, 2021.

- a subset of the IARPA Babel Swahili Language Pack[2] (conversational and scripted telephone speech that spoken in the Nairobi dialect region of Kenya).

The total size of the training corpus is about 74 hours. In our preliminary experiments, we also tried to include a swa dataset (5k instances of Coastal Swahili), provided by the IWSLT-2021 organizers, into the training corpus, but this does not improve the ASR performance. Hence, for the submitted system and for the results reported in the paper, this corpus was not used.

2.1.2 Architecture

In this work, we investigated the impact of using self-supervised learning (Baevski et al., 2020) on the hybrid ASR HMM/DNN acoustic models, as well as on the pipeline ASR+MT system performance. Self-supervised learning (SSL) has shown to be effective for various speech-related tasks including ASR and MT (Schneider et al., 2019; Baevski et al., 2020; Evain et al., 2021; Nguyen et al., 2020) and could be especially beneficial for a low-resource scenario.

We trained several acoustic models (AM) with two different types of input features for comparison: (1) 40-dimensional high-resolution (*hires*) MFCC features; and (2) wav2vec 2.0 features (Baevski et al., 2020) extracted by the multilingual large model pretrained pretrained in 53 languages, XLSR-53-*large* (Conneau et al., 2020).

The phoneme set and transcriptions were the same as in the work (Gelas et al., 2012).

The AMs are state-of-the-art factorized time delay neural networks (TDNN-F) (Povey et al., 2018; Peddinti et al., 2015) and were trained using the Kaldi toolkit (Povey et al., 2011). The models have similar topology (except for the input features): 12 TDNN-F layers (1,024-dimensional, with projection dimension of 128) and a 2232-dimensional output layer. The AMs were trained using lattice-free maximum mutual information (LF-MMI) (Povey et al., 2016) and cross-entropy criteria. Speed and volume perturbations have been applied for data augmentation. 100-dimensional speaker i-vectors were appended to the input features.

We used a 3-gram LM with a 466K vocabulary provided in the ALLFA recipe (Gelas et al., 2012)[3].

[2]IARPA-babel202b-v1.0d, https://catalog.ldc.upenn.edu/LDC2017S05

[3]https://github.com/getalp/ALFFA_PUBLIC

2.2 Neural machine translation system

In order to translate the ASR outputs from source languages to target languages, two neural machine translation systems were built.

2.2.1 Data

For the swa-eng sentence pairs, training dataset for machine translation system includes:

- OPUS[4] english-swahili parallel data : CCAligned and MultiCCAligned (El-Kishky et al., 2020), WikiMatrix, Wikimedia, XLEnt and ParaCrawl.
- 5k parallel swa-eng dataset provided by IWSLT-2021.

The total size of the training dataset for swa-eng is about 3.2M sentence pairs. We applied language identification filtering LangID (Lui and Baldwin, 2012) keeping only swa-eng sentence pairs with correct English. Sentence pairs where the English side is detected as noisy are removed from the swa-eng training dataset. In total, we filter out about 30% of the original training set and obtains a dataset of 2.2M sentence pairs. As for swc-fra NMT system, training data includes parallel corpora made available by the organizers in addition to the available corpora for this language pair on OPUS website. Overall we used a training set of 1.1 M sentence pairs.

2.2.2 Architecture

We propose an NMT model using long short-term memory neural networks (LSTMs) (Hochreiter and Schmidhuber, 1997). NMT systems for swa-eng and for swc-fra were trained using the lstm_luong_wmt_en_de model template, a standard LSTM Encoder-Decoder architecture with Luong-style attention (Luong et al., 2015). Swa-eng system was built at the subword level using a joint BPE vocabulary of 32768 BPE unit, trained using source and target language. Swc-fra NMT model, on its side, was trained at the word level.

2.3 Results

The ASR results in terms of word error rate (WER) are reported in Table 1 on the development datasets for different types of acoustic features. We can see that using wav2vec features significantly decreases the WER and provides about 8% of relative WER reduction for both datasets. Table 2 shows

[4]https://opus.nlpl.eu

the official ASR results on the test datatests. For our submissions, we used wav2vec features only. These ASR results are the best ones among the ASR results submitted by the participants to this task.

Features	Dev swa	Dev swc
MFCC hires	19.90	29.57
wav2vec 2.0	18.34	27.29

Table 1: ASR performance (WER,%) for the development datasets of the low-resource task.

Features	Test swa	Test swc
wav2vec 2.0	31.25	36.75

Table 2: Official ASR performance (WER,%) for the test datasets of the low-resource task.

The MT results in terms of BLEU (Papineni et al., 2002) score are reported in Table 3. Notice that while the WER of the outputs of the ASR fed by wav2vec features is lower than the one fed by MFCC features, for the swc-fra language pair, the BLEU score of the translation from the MFCC-based ASR system is higher than the one got on the wav2vec-based ASR. By lack of time, we did not yet investigate the reason of this, but we will do as soon as possible.

Features	Dev swa-eng	Dev swc-fra
MFCC hires	13.39	9.60
wav2vec 2.0	14.19	9.45
reference text	18.36	14.07

Table 3: MT performance (BLEU) for the development datasets of the low-resource task.

3 Multilingual speech translation

Speech-to-text translation (ST) consists in translating a speech utterance in a source language to a text in another target language (*e.g.*, English audio to French text). In this section, we describe a *multilingual* ST system that can translate from multiple source languages to multiple target ones.

3.1 Data

The data provided for the multilingual ST task is a subset of the Multilingual TEDx corpus (Salesky et al., 2021), in which there are four source languages (Spanish (es), French (fr), Portuguese (pt), and Italian (it)) and five target languages (the aforementioned source languages plus English (en)). The sizes of the ASR talks range from 107 hours (it) to 189 hours (es). Translation data is part of the ASR talks for a given source language. Our experiments were performed in the *constrained* setting where only the provided data for the task is used.

Features	Test swa-eng	Test swc-fra
wav2vec 2.0	12.9	9.1

Table 4: Official MT performance (BLEU) on the test datasets of the low-resource task for the submitted system.

3.2 Model architecture

Our system is based on the Dual-decoder Transformer (Le et al., 2020) which consists of an encoder and two decoders. This architecture jointly transcribes and translates an input speech. Each of the decoders is responsible for one task (ASR or ST) while interacting with each other. We refer the reader to the paper for further details.

We initially followed Le et al. (2020) and used 12 encoder layers, 6 decoder layers, and a hidden dimension of $d = 256$. However, this model produced poor results. We hypothesize that with this configuration, the model capacity is too large for the dataset described in the previous section. In the end, we ended up using only 6 encoder layers and 3 decoder layers (with the same $d = 256$). In addition, we also trained a Transformer model having the same encoder of 6 layers but with only one decoder as the baseline (hereafter called single-decoder model).

3.3 Implementation details

For text pre-processing, we normalize the punctuation and build the vocabulary on the concatenation of the transcript and translation text using SentencePiece (Kudo and Richardson, 2018) without pre-tokenization. We used 10k unigram vocabulary as it performed slightly better than a vocabulary of 8k tokens in our preliminary experiments. The speech features are 80-dimensional log Mel filterbank. Utterances having more than 3000 frames are removed for GPU efficiency. We used SpecAugment (Park et al., 2019) with Librispeech double (LD) policy for data augmentation.

	es-en	es-fr	es-pt	es-it	fr-en	fr-es	fr-pt	pt-en	pt-es	it-en	it-es
Training data (hours)	69	11	42	11	50	38	25	59	-	-	-
Results on the dev sets											
Single decoder	11.56	4.95	19.14	17.98	13.27	12.99	12.21	13.54	8.31	3.88	4.54
Single decoder*	12.75	5.32	21.95	16.82	11.27	12.15	11.01	12.37	10.25	2.24	2.50
Dual-decoder*	18.59	8.02	25.38	19.22	17.81	17.79	15.20	17.63	3.00	4.09	4.81
Official results on the hidden test sets											
Dual-decoder*	20.20	8.20	25.60	11.10	14.40	15.00	14.90	13.20	3.00	4.20	4.60

Table 5: **BLEU on the dev and hidden test sets**. * denotes the use of the transcripts in training.

For the target-forcing mechanism, we prepended a language-specific token to the target sequence (Inaguma et al., 2019; Le et al., 2020). In order to provide good initialization for our multilingual ST system, we separately trained a multilingual ASR system and a multilingual MT one on the allowed data. We then used the weights from the pre-trained ASR encoder, ASR decoder and MT decoder to initialize our ST encoder, ASR decoder, and ST decoder, respectively. We also used the obtained multilingual MT model to augment the training data by translating the transcripts to the target languages as well as translating the translations back to the source languages.

Our model was trained for 150 epochs using the Adam optimizer (Kingma and Ba, 2015) with the inverse square root scheduler. We averaged the last 10 checkpoints and used beam search with a beam size of 5 for decoding. The results reported are detokenized case-sensitive BLEU (Papineni et al., 2002). Our implementation is based on the FAIRSEQ S2T toolkit (Wang et al., 2020).

3.4 Results

Table 5 displays the results on the dev and hidden test sets. One can observe that the Dual-decoder Transformer outperforms the baselines of single decoder on all language pairs except for the pt-es direction where it is surpassed by the single-decoder models. The use of transcripts as additional languages (Gangi et al., 2019) in the single-decoder model improves the results for 4 out of 11 language pairs. Since we aim to obtain a single end-to-end multilingual ST system that can perform many-to-many translation, we selected the Dual-decoder Transformer for our final submission.

4 Conclusion

This paper described the ON-TRAC consortium submissions to the low-resource translation task and to the multilingual speech translation task. Our unique ASR system for both Swahili and Congolese Swahili languages uses XLSR-53 wav2vec features as speech representation input. It got the best results on both Swahili languages (respectively 31.25% and 36.75% of WER). The NMT systems used to translated these transcription into respectively to English and to French got BLEU scores of 12.9 (swa→eng) and 9.1 (swc→fr). The Dual-decoder Transformer we used in the multilingual speech translation got promising results. We did not try a specific strategy to handle language pairs without training data. The low results we got on such language pairs confirm that a specific treatment must be applied in these conditions.

Acknowledgments

This work was funded by the French Research Agency (ANR) through the ON-TRAC project under contract number ANR-18-CE23-0021.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP 2020).

Maha Elbayad, Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Antoine Caubrière, Benjamin Lecouteux, Yannick Estève, and Laurent Besacier. 2020. On-trac consortium for end-to-end and simultaneous speech translation challenge tasks at iwslt 2020. *arXiv preprint arXiv:2005.11861*.

Solene Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, et al. 2021. Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech. *arXiv preprint arXiv:2104.11462*.

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 585–592. IEEE.

Hadrien Gelas, Laurent Besacier, and François Pellegrino. 2012. Developments of swahili resources for an automatic speech recognition system. In *Spoken Language Technologies for Under-Resourced Languages*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 570–577. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3520–3533. International Committee on Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Ha Nguyen, Fethi Bougares, N. Tomashenko, Yannick Estève, and Laurent Besacier. 2020. Investigating Self-Supervised Pre-Training for End-to-End Speech Translation. In *Proc. Interspeech 2020*, pages 1466–1470.

Ha Nguyen, Natalia Tomashenko, Marcely Zanon Boito, Antoine Caubrière, Fethi Bougares, Mickael Rouvier, Laurent Besacier, and Yannick Estève. 2019. On-trac consortium end-to-end speech translation systems for the iwslt 2019 shared task. *arXiv preprint arXiv:1910.13689*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. *CoRR*, abs/2102.01757.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 33–39. Association for Computational Linguistics.