FVLLMONTI

Call: **H2020-FETPROACT-2020-01**

Grant Agreement no. **101016776**

Deliverable D5.1 – Pre-trained speech ASR/MT model and use cases - V1

Start date of the project: 1st January 2021
Duration: 50 months
Project Coordinator: Cristell MANEUX - University of Bordeaux
Contact: Cristell MANEUX - cristell.maneux@ims-bordeaux.fr

DOCUMENT CLASSIFICATION

Title	Pre-trained speech ASR/MT models
Deliverable	D5.1
Estimated Delivery	31/08/2021 (M6+2)
Date of Delivery Foreseen	31/08/2021 (M6+2)
Actual Date of Delivery	31/08/2021 (M6+2)
Authors	Jean-Luc Rouas – P1 – UBx, Leïla Ben Letaifa – P1 – UBx, Georgeta Bordea – P1 – UBx
Approver	Giovanni Ansaloni – P5 – EPFL
Work package	WP5
Dissemination	PU
Version	V1.2
Doc ID Code	D5.1_FVLLMONTI_P1-UBX-20210831
Keywords	Speech Recognition, Machine Translation, Speech Translation

DOCUMENT HISTORY

VERSION	PUBLICATION DATE	CHANGE
1.0	2.08.2021	Initial version from UBx (J.L. Rouas, L. Ben Letaifa, G. Bordea)
1.1	6.08.2021	Updates by UBx from partners feedback
1.2	9.08.2021	Minor changes added by EPFL (G. Ansaloni)
1.3	10.08.2021	Document reviewed and approved by project partners

DOCUMENT ABSTRACT

This document presents the achievements of the partners participating to WP5 during the first semester of the FVLLMONTI project. In this period, work has focused on developing Automatic Speech Recognition (ASR) and Machine Translation (MT) systems using state-of-the-art methods, including neural network transformer architecture. Such activities are the first step towards the WP objectives, as they establish a baseline toward run-time and energy-wise optimization strategies that will be explored for the remainder of the project duration.

Achievements can be summarized as follows:
1 - identification of a common framework for speech recognition and translation
2 - choice of speech recognition/translation datasets
3 - identification of the baseline transformer model settings
4 - initial speech recognition/translation experiments
5 - performance evaluation and analysis of baseline implementations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016776.

TABLE OF CONTENT

LIST OF FIGURES AND TABLES

LIST OF ACRONYMS / GLOSSARY

ASR:	Automatic Speech Recognition
BLEU:	BiLingual Evaluation Understudy
BPE:	Byte Pair Encoding
CER:	Character Error Rate
D:	Deliverable
ESTER:	Evaluation of Speech broadcast news Enriched Transcription systems
Europarl-ST:	European parliament Speech Translation
M:	Month of the project
MT:	Machine Translation
MuST-C:	Multilingual Speech Translation Corpus
PU:	Public
SE:	Speech Enhancement
ST:	Speech Translation
TTS:	Text To Speech
VC:	Voice Conversation
WER:	Word Error Rate
WP:	Work Package

1. Framework and Databases

I. COMMON FRAMEWORK FOR SPEECH RECOGNITION / MACHINE TRANSLATION

The project is proceeding towards the creation of a unified deep learning framework for both automatic speech recognition (or speech to text) and machine translation. Such strategic choice is driven by the expertise of the UBx project partner, and will allow to assess performance/cost benefits of the end-to-end application chain.

We adopt ESPNET as a starting point[1]. The ESPNET project was initiated in December 2017 to mainly deal with end-to-end speech recognition experiments based on sequence-to-sequence modeling. The project has grown rapidly and now covers a wide range of speech processing applications. Now ESPNET also includes Text To Speech (TTS), Voice Conversation (VC), Speech Translation (ST), and Speech Enhancement (SE) with support for beamforming, speech separation, denoising, and dereverberation. All applications are trained in an end-to-end manner, thanks to the generic sequence-to-sequence modeling properties, and they can be further integrated and jointly optimized. Also, ESPNET provides reproducible all-in-one recipes for these applications with state-of-the-art performance in various benchmarks by incorporating transformer, advanced data augmentation, and conformer models.

ESPNET is Licensed under the Apache License, Version 2.0.

The project partner UBx has installed the ESPNET framework on a variety of hardware / software hosts:

- CPUs: Intel(R) Core(TM) i7-6950X CPU @ 3.00GHz/ Intel(R) Core(TM) i9-7940X CPU @ 3.10GHz / Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz / Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz
- RAM: 132Go / 192 Go / 396 Go
- GPUs: NVIDIA TITAN X / NVIDIA GeForce RTX 2080 Ti / Quadro RTX 8000 / Quadro RTX 6000
- Operating systems: Ubuntu 16.04 LTS / Ubuntu 18.04 LTS / Ubuntu 20.04 LTS

II. AUTOMATIC SPEECH RECOGNITION DATASETS

For automatic speech recognition, we are currently considering databases in the French and English languages:

- For English, experiments target LIBRISPEECH database[2] . LIBRISPEECH is a corpus of approximately 1000 hours of 16kHz read English speech, prepared by Vassil Panayotov with the assistance of Daniel Povey [1]. The data is derived from read audiobooks from the LibriVox project[3], and has been carefully segmented and aligned. The size of the corpus makes it impractical to distribute it as a single large archive. Thus, the training portion of the corpus is split into three subsets, according to recording quality. The development and test sets have been split in two parts (see Table 1).

[1] https://espnet.github.io/espnet/index.html
[2] https://www.openslr.org/12
[3] https://librivox.org/

subset	hours	Speakers
dev-clean	5.4	40
test-clean	5.4	40
dev-other	5.3	33
test-other	5.1	33
train-clean-100	100.6	251
train-clean-360	363.6	921
train-clean-500	496.7	1166

Table 1: Data subsets in LIBRISPEECH

- For French, we target the ESTER database[4] The ESTER corpus was produced within the French national project ESTER (Evaluation of Speech broadcast news Enriched Transcription systems), as part of the Technolangue programme funded by the French Ministry of Research and New Technologies (MRNT). The ESTER project enabled to carry out a campaign for the evaluation of Broadcast News enriched transcription systems using French data. The dataset contains about 245 hours of orthographically transcribed news broadcast, including annotations of named entities. The ESTER campaign was conducted in two phases: ESTER 1 and ESTER 2 [2]. Each phase produced about 100 hours of transcribed speech. An additional set of 45 hours extracted from the not transcribed portion of the ESTER 1 database was annotated as part of the EPAC project[5]. Apart from the training data, development and test sets, each containing 6 hours of radio news broadcast, was provided during ESTER 2.

Subset	hours
dev	5.46
test	6.34
train	231

Table 2: Data subsets in ESTER

[4] http://catalogue.elra.info/en-us/repository/browse/ELRA-S0338/
[5] http://epac.univ-lemans.fr/

III. MACHINE TRANSLATION DATASETS

We identified the following machine translation datasets of interest for the FVLLMONTI project:

- Europarl-ST[6] [10]. Europarl-ST is a Multilingual Speech Translation Corpus, that contains paired audio-text samples for Speech Translation, constructed using the debates carried out in the European Parliament in the period between 2008 and 2012. Table 3 gives an overview of the datasets that will be used in our project.

Source/Target	EN	FR	DE	IT
EN	-	81	83	80
FR	32	-	21	20
DE	30	18	-	17
IT	37	21	21	-

Table 3: Size of training data subsets used from Europarl-ST (in hours)

- MuST-C[7] [8]: MuST-C is a multilingual speech translation corpus whose size and quality facilitates the training of end-to-end systems for speech translation from English into several languages. For each target language, MuST-C comprises several hundred hours of audio recordings from English TED Talks, which are automatically aligned at the sentence level with their manual transcriptions and translations. In the table below we give some statistics about the datasets of interest for our work.

Subset	Size	EN-DE	EN-FR	EN-IT
dev	**talks**	11	11	11
	sentences	32	1412	1309
	source words	26686	26754	25972
	target words	25613	27363	24154
	time	2h 32m 44s	4h 08m 54s	2h 29m 11s
train	**talks**	2043	2460	2324
	sentences	229703	275085	253588
	source words	4196735	5067070	4772289
	target words	3869794	5163708	4421764

[6] https://www.mllp.upv.es/europarl-st/
[7] https://ict.fbk.eu/must-c/

	time	400h 02m 28	484h 31m 21s	457h 19m 52
tst-COMMON	**talks**	27	27	27
	sentences	2641	2632	2574
	source words	46082	46127	46025
	target words	43775	48174	41360
	time	4h 08m 54s	4h 09m 27s	4h 09m 41s

Table 4: Data subsets used from MuST-C

- LIBRITRANS[8] [6]: This corpus is an augmentation of the LIBRISPEECH ASR corpus(1000h) [1] and contains English utterances (from audiobooks) automatically aligned with French text. The dataset is currently available for download[9].This dataset offers ~236h of speech aligned to translated text that is split as can be seen in Table 4. The 100h subset was specifically designed for direct speech translation training and evaluation [5]. In this subset, the best 100 hours according to cross language alignment scores were extracted. The test and development sets correspond also to the highest rated alignments, composed of clean speech segments only. The remaining data (extended train) is more noisy, as it contains more incorrect alignments.

subset	**hours**
dev	2
test	3,44
train	100
extended train	130

Table 5: Data subsets in LIBRITRANS

- A limitation of the datasets described previously for the machine translation task is that they cover monologue data rather than dialogue. To also target dialogues in our experiments, we plan to employ the data collected during the COMPRISE project and made available by the COMPRISE consortium[10].

[8] https://github.com/alicank/Translation-Augmented-LibriSpeech-Corpus
[9] https://perscido.univ-grenoble-alpes.fr/datasets/DS91.
[10] https://www.compriseh2020.eu/

2. Parameters and Settings

Transformer-type architecture has seen significant interest recently, due to its good performance in various sequence-to-sequence applications, such as MT [11], language modelling and ASR [12]. Hence the choice of these models for the development of our machine recognition and translation systems. The architecture and optimization parameters are described below.

I. SPEECH RECOGNITION

For speech recognition, we use 4 different models, two for English and two for French. The two English models are built on the LIBRISPEECH corpus and on the LIBRITRANS corpus. We name them LIBRISPEECH-Transformer models and LIBRITRANS-Transformer models, respectively, in the rest of this document. The two French models are built on the ESTER corpus and are named ESTER-Transformer and ESTER-Conformer.

For model training, data pre-processing is necessary. This consists of data preparation and parameter extraction. The data preparation step consists of cleaning the transcript and putting the audio/text in a predefined form[11].

Regarding the extraction of the parameters, we proceeded in several steps, performing:

- Speed perturbation to increase the training data amount. This method stretches or squeezes the duration of audio signals by changing the sample rate of audio waveforms. Three speed coefficients were applied, leading to an overall data volume of approximately 2800h for LIBRISPEECH, 300h for LIBRITRANS and 700h for ESTER.
- Extraction of characteristic coefficients. For ESTER and LIBRITRANS, these are 80 fbank coefficients + pitch.
- Deletion of overly long and short sentences. A sentence is considered too long if the audio signal has more than 3000 frames or transcription contains more than 400 characters. It is considered too short if the audio is less than 10 frames long or if its transcription contains no characters.
- Normalization of parameters with respect to mean and variance.
- Definition of the vocabulary. For the French experiments, the vocabulary is mainly formed by the French characters. For English, the Byte Pair Encoding (BPE) subword segmentation algorithm was used to extract 5000 and 1000 BPEs for LIBRISPEECH and LIBRITRANS respectively.
- For ESTER and LIBRITRANS, an additional data augmentation using the SpectAugment [14] method was applied before model training. This method directly processes spectrograms rather than waveforms, as opposed to speed perturbation.

A. English models

For English, we took the advantages of existing models trained on the LIBRISPEECH database, available in the ESPNET ZOO repository[12] . A detailed description of these models is provided in [3][7].

[11] https://kaldi-asr.org/doc/data_prep.html
[12] https://github.com/espnet/espnet_model_zoo

As experiments for machine translation use the LIBRITRANS dataset, for comparison we also built our own transformer models from scratch using the LIBRITRANS training dataset, using the train-clean-100 subset of LIBRISPEECH. The specifications of these models are the following:

LIBRISPEECH-Transformer model (pre-trained)	***LIBRITRANS-Transformer model (own)***
# minibatch related batch-size: 20 batch_bins: 15000000 # optimization related criterion: loss early-stop-criterion:"validation/main/loss" sortagrad: 0 opt: adam lr: 0.002 optimizer-warmup-steps: 25000 epochs: 100 patience: 0 accum-grad: 6 grad-clip: 5.0 # hybrid CTC/attention ctc_weight: 0.3 # network architecture ## encoder related etype: custom custom-enc-input-layer: enc-block-arch: - type: transformer d_hidden: 512 d_ff: 2048 heads: 8 dropout-rate: 0.1 att-dropout-rate: 0.1 enc-block-repeat: 18 ## decoder related dec-block-arch: - type: transformer d_hidden: 512 d_ff: 2048 heads: 8 dropout-rate: 0.1 att-dropout-rate: 0.1 dec-block-repeat: 6	# minibatch related batch-size: 64 maxlen-in: 512 maxlen-out: 150 # optimization related sortagrad: 0 opt: noam accum-grad: 2 grad-clip: 5 patience: 0 epochs: 100 dropout-rate: 0.1 # hybrid CTC/attention mtlalpha: 0.3 # network architecture # encoder related elayers: 12 eunits: 2048 # decoder related dlayers: 6 dunits: 2048 # attention related adim: 256 aheads: 4 # transformer specific setting transformer-input-layer: conv2d transformer-lr: 5.0 transformer-warmup-steps: 25000 transformer-attn-dropout-rate: 0.0

The disk space occupied by these LIBRISPEECH-Transformer and LIBRITRANS-Transformer models is 380MB and 107MB respectively.

B. French models

For the French language, pre-trained models are unfortunately not available. We thus trained models ourselves, building them on the ESTER database. Two models have been derived: one based on a transformer architecture and the other on conformer architecture. Their specifications are as follows:

ESTER-Transformer model	*ESTER-Conformer model*
# minibatch related batch-size: 64 maxlen-in: 512 maxlen-out: 150 # optimization related criterion: loss early-stop-criterion: “validation/main/loss” sortagrad: 0 opt: noam noam-lr: 3.0 noam-adim: 512 optimizer-warmup-steps: 25000 epochs: 100 patience: 0 accum-grad: 2 grad-clip: 5.0 # network architecture ## encoder related etype: custom custom-enc-input-layer: vgg2l enc-block-arch: - type: transformer d_hidden: 512 d_ff: 2048 heads: 4 dropout-rate: 0.1 att-dropout-rate: 0.1 enc-block-repeat: 18 ## decoder related dec-block-arch: - type: transformer d_hidden: 512 d_ff: 2048 heads: 4 dropout-rate: 0.1 att-dropout-rate: 0.1 dec-block-repeat: 6 ## joint network related joint-dim: 512	# minibatch related batch-size: 64 maxlen-in: 512 maxlen-out: 150 # optimization related criterion: loss early-stop-criterion: “validation/main/loss” sortagrad: 0 opt: noam noam-lr: 1.0 noam-adim: 512 optimizer-warmup-steps: 25000 epochs: 100 patience: 0 accum-grad: 2 grad-clip: 5.0 # network architecture ## general custom-enc-positional-encoding-type: rel_pos custom-enc-self-attn-type: rel_self_attn custom-enc-pw-activation-type: swish ## encoder related etype: custom custom-enc-input-layer: vgg2l enc-block-arch: - type: conformer d_hidden: 512 d_ff: 2048 heads: 4 macaron_style: True use_conv_mod: True conv_mod_kernel: 15 dropout-rate: 0.3 att-dropout-rate: 0.3 enc-block-repeat: 12 ## decoder related dtype: lstm dlayers: 1 dec-embed-dim: 1024 dunits: 512 dropout-rate-embed-decoder: 0.2

	dropout-rate-decoder: 0.1 ## joint network related joint-dim: 512

The disk space occupied by the ESTER-Transformer and ESTER-Conformer models is 228MB and 310MB, respectively.

II. MACHINE TRANSLATION

The configuration of pre-trained models for machine translation is as follows:

MuST-C model	*LIBRITRANS model*
# minibatch related batch-size: 96 maxlen-in: 100 maxlen-out: 100 # optimization related sortagrad: 0 opt: noam accum-grad: 1 grad-clip: 5 patience: 0 epochs: 100 dropout-rate: 0.1 # network architecture # encoder related elayers: 6 eunits: 2048 # decoder related dlayers: 6 dunits: 2048 # attention related adim: 256 aheads: 4 tie-src-tgt-embedding: false tie-classifier: false # label smoothing lsm-weight: 0.1 # transformer specific setting backend: pytorch model-module: "espnet.nets.pytorch_backend.e2e_mt_transformer :E2E" transformer-lr: 1.0 transformer-warmup-steps: 8000 transformer-attn-dropout-rate: 0.1	# minibatch related batch-size: 96 maxlen-in: 100 maxlen-out: 100 # optimization related sortagrad: 0 opt: noam accum-grad: 1 grad-clip: 5 patience: 0 epochs: 100 dropout-rate: 0.1 # network architecture # encoder related elayers: 6 eunits: 2048 # decoder related dlayers: 6 dunits: 2048 # attention related adim: 256 aheads: 4 tie-src-tgt-embedding: false tie-classifier: false # label smoothing lsm-weight: 0.1 # transformer specific setting backend: pytorch model-module: "espnet.nets.pytorch_backend.e2e_mt_t ransformer:E2E" transformer-lr: 1.0 transformer-warmup-steps: 8000 transformer-attn-dropout-rate: 0.1

transformer-length-normalized-loss: false transformer-init: xavier_uniform # pre-training related enc-init-mods: encoder.embed,encoder.encoders,encoder.after_norm dec-init-mods: decoder.embed,decoder.decoders,decoder.after_norm,decoder.output_layer	transformer-length-normalized-loss: false transformer-init: xavier_uniform # pre-training related enc-init-mods: encoder.embed,encoder.encoders,encoder.after_norm dec-init-mods: decoder.embed,decoder.decoders,decoder.after_norm,decoder.output_layer

The disk space needed to store the MT models is about 90MB for the MuST-C dataset and about 70MB for the LIBRITRANS model.

3. PERFORMANCE EVALUATION

Evaluation of the identified models is ongoing. We are proceeding in three stages:

1. Performances of the speech transcription systems alone, evaluated on speech recognition databases (LIBRISPEECH, ESTER).
2. Performances of the machine translation systems, on machine translation datasets (Must-C, LIBRITRANS).
3. Performances of the whole chain, speech recognition followed by machine translation, on a specific translation database that includes speech samples (LIBRITRANS). The performances of the the speech recognition systems have been, as of now, evaluated on the ASR specific database. The evaluation of the performance of the translation system using the outputs of the speech recognition systems is planned for the upcoming months.

I. SPEECH RECOGNITION

The standard metric of ASR evaluation is the Word Error Rate (WER). It is defined as the proportion of word errors to words processed:

$$WER = Sub + Del + Ins / N$$

where Sub is the number of substitutions, Del is the number of deletions, Ins is the number of insertions and N is the number of words in the reference transcription. Character Error Rate (CER) is computed in the same way. Both of them are provided to evaluate the different ASR systems.

A. English

In Table 6 and Table 7 below, results using the LIBRISPEECH-Transformer system on both test parts of the LIBRISPEECH corpus are described. These results are better than those reported in recent publications using ESPNET and Transformers on the same database [3]. It is to be noted that performances are expected to be slightly worse for the test-other subset since it contains speech from non-native English speakers.

Test_clean	Snt	Wrd	Corr	Sub	del	Ins	**Err**	S.Err
WER	2620	52576	97.6	2.0	0.3	0.3	**2.6**	30.7
CER	2620	281530	99.2	0.3	0.4	0.3	**1.0**	30.7

Table 6: Results on LIBRISPEECH test-clean

Test_other	Snt	Wrd	Corr	Sub	del	Ins	**Err**	S.Err
WER	2639	52343	94.5	4.7	0.8	0.7	**6.2**	49.6
CER	2939	272758	97.7	1.1	1.2	0.7	**3.0**	49.6

Table 7: Results on LIBRISPEECH test-other

B. French

The performances obtained by the ESTER-Transformer system are given in Table 8. These results are similar to the best results obtained on the ESTER database mentioned in [4], although these were not obtained using an end-to-end system.

Test	Snt	Wrd	Corr	Sub	del	Ins	**Err**	S.Err
WER	6193	74016	87.2	10.6	2.2	2.2	**14.9**	65.2
CER	6193	424782	95.6	2.0	2.4	2.1	**6.5**	66.7

Table 8: Results for the ESTER-Transformer system on the ESTER database (French)

The results for the ESTER-Conformer system are given in Table 9. These results are significantly higher than those obtained using the transformer-based system.

Test	Snt	Wrd	Corr	Sub	del	Ins	**Err**	S.Err
WER	6193	74016	89.8	8.3	1.9	1.6	**11.8**	59.2
CER	6193	424782	96.6	1.4	1.9	1.6	**5.0**	59.2

Table 9: Results for ESTER-Conformer on the ESTER database (French)

II. MACHINE TRANSLATION

The most widely recognised evaluation metric for evaluating the quality of text which has been machine-translated is the BLEU score [13]. Scores are calculated for individual translated segments by comparing them with a set of human-produced reference translations and then averaged for the whole corpus. The BLEU score indicates how similar the candidate text is to the reference texts, with higher values representing more similar texts. The performances obtained by the Transformer-based pre-trained models on the MuST-C database are provided in Table 10 and for the LIBRITRANS database in Table 11.

Test	En->De	En->Fr	En->It
Main direction	30.16	43.02	31.08
Reverse direction	35.37	43.70	33.39

Table 10: Results for Transformer-based MT on the MuST-C database (4-gram BLEU)

Regardless of the direction of translation the best results are obtained on this dataset for translating between English and French.

Test	BLEU	1-gram	2-gram	3-gram	4-gram	BP	ratio	hyp_len	ref_len
Transformer	18.0	50.0	23.8	13.1	7.4	0.982	0.982	43115	43904
RNN	18.39	52.6	25.6	14.3	8.1	0.926	0.929	40950	44080

Table 11: Results for Transformer-based MT and RNN baseline on the LIBRITRANS database

Across languages the results are considerably better for the MuST-C dataset compared to the LIBRITRANS dataset. This can partially be explained by the larger size of the MuST-C dataset but mainly it is due to the different complexity of the tasks involved, that is translating literary text compared to translating speeches intended for a broad audience.

III. PERFORMANCES ON THE COMPLETE TRANSLATION CHAIN

The complete translation chain (called Cascade-ST) as can be seen in Figure 1. Its evaluation is being carried out in two steps:

- Evaluation of the speech transcription systems on the dedicated speech-to-translation database (LIBRITRANS)
- Evaluation of the machine translation systems using the outputs of the speech recognition systems. This part of the evaluation is still ongoing.

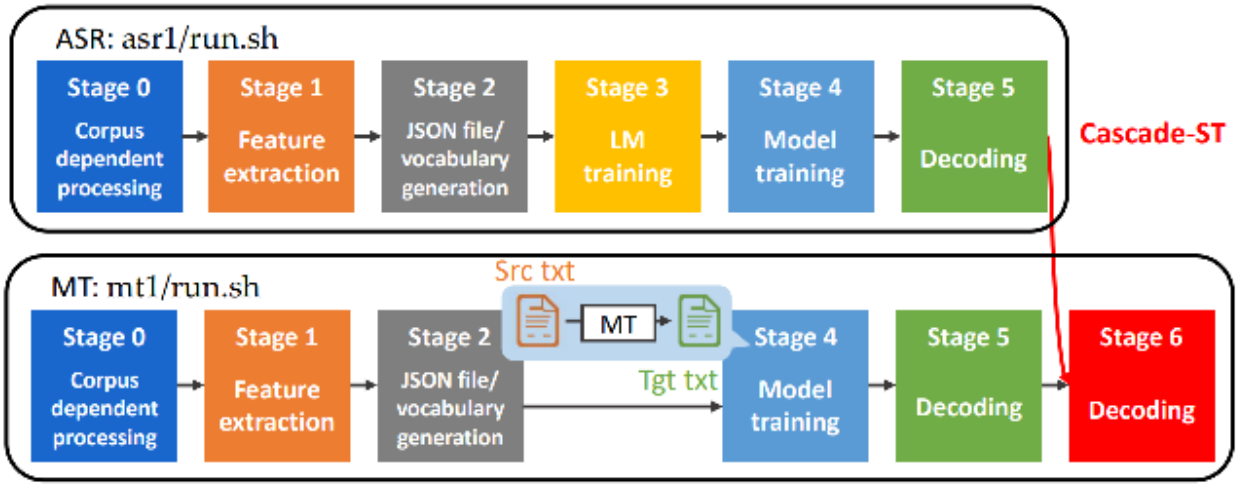

Figure 1: Overview of the complete ASR/MT chain in ESPNET (figure adapted from [9])

The speech recognition step is the first step in the speech translation system. Since the databases used for machine translation are not the same as the databases used for speech recognition, we have to evaluate the performances on these databases to ensure suitable performance of the whole process. It is also to be noted that even though the databases used for machine translation include speech data and the associated orthographic transcriptions, the amount of data provided by these databases may not be sufficient for proper training of deep neural networks models for speech recognition. Hence the need for external databases for speech recognition systems design and training, that need further testing on the samples of the machine translation databases.

Two sets of experiments have been conducted using the LIBRITRANS database, using Transformer models learned from the LIBRITRANS and LIBRISPEECH datasets, named LIBRITRANS-Transformer and LIBRISPEECH-Transformer in the following.

The results obtained using LIBRITRANS-Transformer models (Table 12) are above the state-of-the-art results reported in [5] (15.1% WER).

	Snt	Wrd	Corr	Sub	Del	Ins	Err	S.Err
WER	2048	36336	94.8	4.7	0.5	0.8	**6.1**	45.2
CER	2048	60986	94.7	3.9	1.3	0.6	**5.9**	45.7

Table 12: Transcription results for LIBRITRANS using LIBRITRANS-Transformer models

Still, even better performances are obtained when using the LIBRISPEECH-Transformer models (Table 13), achieving a 2.0% WER.

	Snt	Wrd	Corr	Sub	Del	Ins	Err	S.Err
WER	2048	36336	98.2	1.4	0.3	0.2	**2.0**	22.7
CER	2048	193184	99.5	0.1	0.4	0.4	**0.9**	22.7

Table 13: LIBRITRANS transcription results with LIBRISPEECH-Transformer models

4. CONCLUSION

This document describes the initial work carried out in WP5 of the FVLLMONTI project concerning the development of Pre- speech ASR/MT models. Our achievements so far are summarized as follows:

(i) identification of a common framework (ESPNET) for both speech recognition and translation

(ii) selection of appropriate existing datasets for Speech Recognition (LIBRISPEECH for English, ESTER for French) and for Machine Translation (MuST-C and LIBRITRANS). Other databases may also be studied in the future, depending on the requirements of the project.

(iii) Finally, we proposed baseline settings for the models (either using pre-trained available models or in-house models) and evaluated their performances. Evaluation results show that we can reach or even surpass state-of-the-art performances.

Work is ongoing in assessing the performance of the complete ASR/MT chain. The outcome of this study will be the springboard for the optimization effort to be undertaken in the rest of the project.

REFERENCES

[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, 'Librispeech: An ASR corpus based on public domain audio books', in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Queensland, Australia, Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.

[2] S. Galliano, G. Gravier, and L. Chaubard, 'The ester 2 evaluation campaign for the rich transcription of french radio broadcasts', 2009.

[3] S. Karita et al., 'A Comparative Study on Transformer vs RNN in Speech Applications', in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Dec. 2019, pp. 449–456. doi: 10.1109/ASRU46091.2019.9003750.

[4] F. Boyer and J.-L. Rouas, 'End-to-End Speech Recognition: A review for the French Language', Oct. 2019. Accessed: Feb. 03, 2021. [Online]. Available: http://arxiv.org/abs/1910.08502

[5] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, 'End-to-End Automatic Speech Translation of Audiobooks', Calgary, Alberta, Canada, Apr. 2018. Accessed: Jul. 29, 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01709586

[6] Ali Kocabiyikoglu, Laurent Besacier, Olivier Kraif. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. LREC (Language Resources and Evaluation Conference), Feb 2018, Miyazaki, Japan.

[7] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang and Yuekai Zhang. Recent Developments on ESPnet Toolkit Boosted by Conformer. CoRR abs/2010.13956 (2020).

[8] Cattoni, Roldano, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. "MuST-C: A multilingual corpus for end-to-end speech translation." Computer Speech & Language 66 (2021): 101155.

[9] Inaguma, H., Kiyono, S., Duh, K., Karita, S., Yalta, N., Hayashi, T. and Watanabe, S., 2020, July. ESPnet-ST: All-in-One Speech Translation Toolkit. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 302-311).

[10] Iranzo-Sánchez, Javier, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. "Europarl-ST: A multilingual corpus for speech translation of parliamentary debates." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8229-8233. IEEE, 2020.

[11] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin. Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[12] Linhao Dong; Shuang Xu and Bo Xu. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. ICCASP 2008.

[13] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318. 2002.

[14] Bahar P, Wilken P, Alkhouli T, Guta A, Golik P, Matusov E, Herold C. "Start-before-end and end-to-end: Neural speech translation by apptek and rwth aachen university." InProceedings of the 17th International Conference on Spoken Language Translation 2020 (pp. 44-54).