Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D3.7 Final report on speech and natural language processing

Work Package	3
Responsible Partner	LIA
Author(s)	Yannick Estève
Contributors	Antoine Caubrière, Gaëlle Laperrière, Guntis Barzdins, Jarod Duret, Jean-François Bonastre, Marcely Zanon Boito, Natalia Tomashenko, Salima Mdhaffar, Titouan Parcollet, Roberts Dargis, Yannick Estève, Raul Monteiro
Reviewer	Christoph Schmidt
Version	1.0
Contractual Date	31 March 2024
Delivery Date	28 March 2024
Dissemination Level	Public

Version History

Version	Date	Description
0.1	03/03/2024	ToC updated
0.2	10/03/2024	Partner content added
0.3	13/03/2024	Internal review ready
1.0	26/03/2024	Publishable version

Executive Summary

This report presents the progress made during the SELMA project on speech and language processing. For speech processing, the research work focused mainly on the use of end-to-end neural models, especially based on model pretrained under self-supervision and on the use of some very recent evolutions of wav2vec 2.0 models like SAMU-XLSR (Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation) for cross lingual transfer. Discrete speech units have also been investigated, mainly for textless speech-to-speech translation. Whisper models have also been considered.

> For speech processing, the work focused mainly on the use of foundation models pretrained by self-supervision to address cross-lingual knowledge transfer

The SELMA project was strongly involved in the LeBenchmark and LeBenchmark2.0 initiatives that permitted to pretrained wav2vec 2.0 models on 7K hours of speech in French and compare them to wav2vec 2.0 models pretrained on English-only data or multilingual data (containing 53 different languages).

These models have also been fine-tuned on downstream tasks directly related to the SELMA project: speech recognition, speech translation, semantic concept extraction from speech and named entity recognition from speech.

During the first year of the project, some baseline automatic speech recognition systems driven by hybrid Hidden Markov Model and Deep Neural Network (HMM/DNN) acoustic models have been developed for some languages (English, French, Latvian). Some of these ASR systems have been integrated into the SELMA platform as NLP components delivered as Docker containers.

A first speech synthesis engine has been built on Brazilian Portuguese broadcast news provided by Deutsche Welle.

During the second year, we prepared the data to pretrain a SELMA wav2vec 2.0 and offered solutions to deal with low resource scenario for spoken language understanding (SLU) and to port an SLU model from one language to another. We also developed end-to-end solutions dedicated to low resource scenarios for speech translation.

During the last period, we have extended our work to textless speech-to-text translation that preserves the expressivity present in the source utterance, reaching our main results during the third year. We also have proposed the M-PHANTOM model that permits taking into account user feedback to improve the quality of the speech recognition.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

Work Package 3 aimed to develop and make advances in state-of-the-art natural language processing technologies, with a special focus on speech processing. In the last decade, such technologies have made considerable progress through the emergence of the deep learning paradigm, but in many tasks, these approaches are still far from solving the most relevant research questions.

One very current hot topic in the speech and language research community is the use of models pretrained by self-supervision. Such deep neural models are trained on huge amounts of unlabeled data. The BERT model, which is dedicated to text processing, has been introduced by Google (Devlin 2019, https://arxiv.org/abs/1810.04805): the main state-of-art systems for any NLP tasks are based on the use of deep neural models derived from BERT. The use of BERT-like models consists of first pretraining a model through self-supervised learning on a very huge amount of unlabeled data and then fine-tuning it on a smaller amount of in-domain labeled data by supervised learning.

Such an approach has been proposed for speech processing with the introduction of the wav2vec models in 2019 by Facebook (Schneider 2019, https://arxiv.org/abs/1904.05862). Significant improvements were proposed in 2020 with the wav2vec 2.0 models (Baevski 2020, https://arxiv.org/abs/2006.11477): it was shown that it is possible to reach low word error rates (<10%) by exploiting only 10 minutes of manually transcribed speech (audiobooks), after pretraining on 960 hours of untranscribed audio.

Pretraining such a model needs a lot of computation power, and a lot of questions are still open about their robustness to acoustic conditions and languages. In the framework of the SELMA project, we brought strong efforts during this first year to master this approach and to pretrain French wav2vec 2.0 models and fine-tuned them into several downstream task. This work is presented in **Section 2**. This work was made in association with external partners (University of Grenoble-Alpes, France) and was possible thanks to the use of the French Jean Zay supercomputer. Some convincing results are presented in this report in **Section 3**. Leveraging this experience, a SELMA model tailored for Deutsche Welle's multilingual broadcast news audio has been developed: during Y2 we have collected and prepared the training data and trained the model during Y3. This model, the SELMA-19 wav2vec2.0 model, is presented in **Section 4**.

In addition to this study on wav2vec 2.0 models, we have worked on speech synthesis on Deutsche Welle data (from Brazilian Portuguese and Urdu broadcast news): our neural architecture is presented in **Section 5**.

We also built more classical hybrid HMM/DNN ASR systems that have been integrated into the SELMA platform, described in **Section 6**.

We also focused on low-resource languages. We proposed a new approach for low resource scenarios in the context of named entity recognition from speech through an end-to-end neural approach – a scientific publication has been submitted, accepted and presented at Interspeech 2022 (Mdhaffar et al., 2022). We proposed new contributions on language portability of end-to-end models dedicated to semantic extraction from speech – scientific publications have been submitted and accepted in this topic, for instance (Laperrière et al., 2023, Laperrière et al., 2023b). This work is described in **Section 7**. In the context of low-resource languages, we also participated in shared tracks in the IWSLT international challenge dedicated to speech translation for low resource languages, especially for spoken Tamasheq to written French and spoken Tunisian to written English (Laurent et al., 2023). These works are presented in **Section 8**.

In SELMA, we explore a challenging task that aims to process speech-to-speech translation by preserving the expressivity of the source utterances. For instance, the linguistic content of sentences uttered by a speaker who speaks loudly and fast will be translated with a synthetic voice that keeps these characteristics, speaking loud and fast. Our contributions (Duret et al., 2023, Duret et al., 2023b) are based on textless speech-to-speech translation through discrete speech units, emotion recognition, and neural speech synthesis. This work is summarized in **Section 9**.

An important and challenging point in SELMA is about reinjecting user feedback to improve the systems. Our main contribution is the M-Phantom model, presented in **Section 10,** that relies on a combination of keyword spotting and relevant prompting applied to a Whisper model.

2. The *LeBenchmark* initiative

End-to-end speech recognition and translation based on speech unit representation learned through self-supervised training.

Self-Supervised Learning (SSL) based on huge amounts of unlabelled data has been explored successfully for image and natural language processing (Bachman et al., 2019; Chen et al., 2020; Devlin et al., 2018; Raffel et al., 2019). Researchers investigated SSL from speech as well and successfully improved performance on downstream tasks such as speech recognition (Baevski et al., 2019; Kawakami et al., 2020).

As SSL from speech is a rapidly evolving domain, new models are unfortunately evaluated on different datasets, most of which focus on the English language. To carefully assess the progress of speech SSL model-wise and application-wise, common benchmarks are needed. While NLP benchmarking is now widely discussed (Ruder, 2021), multi-task benchmarks are less common in speech even though the field has a long tradition of evaluation (see for instance long-term NIST and DARPA shared tasks for ASR).

In our papers Evain et al., 2021-A and Evain et al., 2021-B, we proposed to contribute to this by providing a reproducible and multifaceted benchmark for evaluating speech SSL models. By *benchmark*, and following the definition of Schlangen, 2021, we mean an ensemble of tasks that allow to discriminate learners (*i.e.,* SSL models) based on their ability to perform well on those tasks.

We propose an initial set of four main tasks (10 sub-tasks overall), measuring specific speech challenges in the French language: Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Speech Translation (AST), and Emotion Recognition (AER). In this document, we present the main results obtained for the first three tasks, while also including results on Named Entity Recognition (NER). The totality of our results can be found in the original papers (Evain et al., 2021-A and Evain et al., 2021-B) as well as on the website's leader board: http://lebenchmark.com .

In summary, our investigation facilitates the evaluation of the impact of pre-trained speech models that differ along several dimensions: language used for pre-training (French, English, multilingual), amount of raw speech used for SSL pre-training (1k, 3k, or 7k hours), model size (base, large). For reproducibility, we also provide pre-trained SSL models learned on a large and heterogeneous collection of speech utterances and believe this is a strong contribution to speech technologies in French.

2.1 Background

SSL has been proposed as an interesting alternative for data representation learning, as it requires no annotated data. Such learned representations have been very successful in computer vision (Bachman et al., 2019; Chen et al., 2020), and language (Devlin et al., 2018, Peters et al., 2018). SSL from speech consists of resolving *pseudo-tasks*, which do not require human annotation, as a pre-training for the real tasks. These *pseudo-tasks* target predicting the next samples or solving ordering problems. For instance, Autoregressive Predictive Coding (APC) considers the sequential structure of speech and predicts information about a future frame (Chung et al., 2019; Chung and Glass, 2020-A), whereas Contrastive Predictive Coding (CPC) distinguishes a future speech frame from distractor samples (Baevski et al., 2019, Schneider et al., 2019), which is an easier learning objective compared to APC. Such representations have been shown to improve performance in several speech tasks (Chung and Glass, 2020-B), while being less sensitive to domain and/or language mismatch (Kawakami et al., 2020) and being transferable to other languages (Riviere et al., 2020).

In 2020, a strong speech SSL baseline appeared: the Wav2Vec2.0 model (Baevski et al., 2020) which relies on the CPC idea of Baevski et al., 2019 and Schneider et al., 2019 but with *discrete* speech units that are used as latent representations and fed to a Transformer network to build contextualized representations. Several other bi-directional encoders were also proposed: Speech-XLNet (Song et al., 2019), Mockingjay (Liu et al., 2019) and Wang et al., 2020. A few recent studies were also related to multilingual SSL models trained on very large multilingual corpora (Conneau et al., 2020, Wang et al., 2021).

While there are multiple evaluation benchmarks to assess pre-trained models in NLP (for instance *lue* for English, *flue* for French, and *klue* for Korean), we are aware of only one similar initiative for speech SSL model evaluation: the Speech processing Universal PERformance Benchmark (SUPERB) (Yang et al., 2021) which however targets English only and does not share pre-trained SSL models as we do.

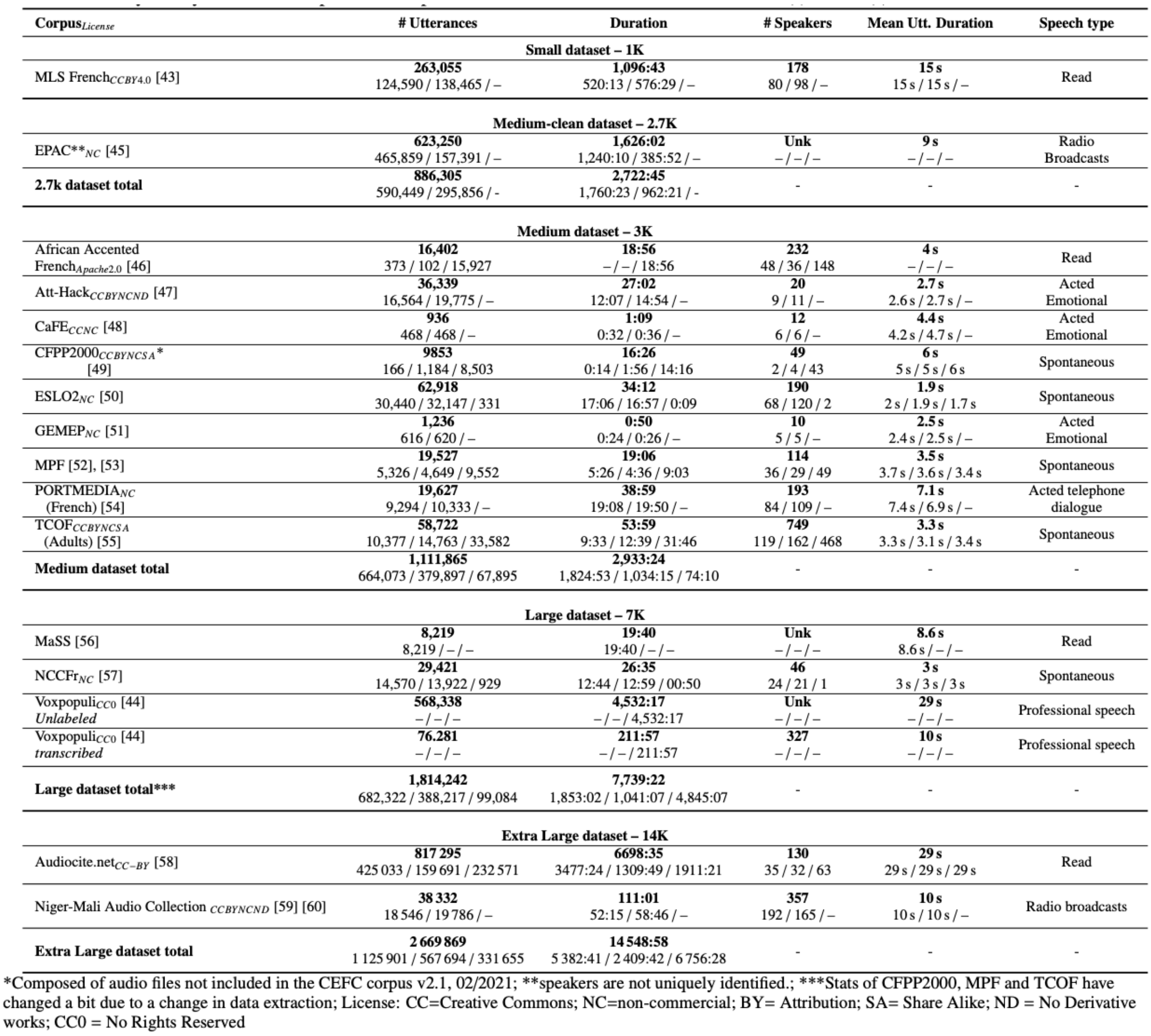
2.2 Gathering a Large and Heterogeneous Speech Collection in French

Large multilingual corpora that include French have been made available, such as MLS (Pratap et al., 2020, 1,096 hours) and Voxpopuli (Wang et al., 2021, +4,500 hours). However, these are restricted to either read or well-prepared speech, failing to provide diversity in the speech samples, such as accented, spontaneous and/or affective speech.

We gathered a large variety of speech corpora in French that cover:

- Different accents: MLS (Pratap et al., 2020), African Accented Speech (SLR57), CaFE (Gournay et al., 2018);
- Acted emotions: GEMEP (Bänziger et al., 2012), CaFE (Gournay et al., 2018), Att-Hack (Le Moine et al., 2020);
- Telephone dialogues : PORTMEDIA (Lefèvre et al., 2012);
- Read sentences: MLS (Pratap et al., 2020), African Accented French (SLR57), MaSS (Boito et al., 2020);
- Spontaneous sentences: CFPP2000 (Branca-Rosoff et al., 2012), ESLO2 (Eshkol-Taravella et al., 2012), MPF (ORTOLANG-MPF), TCOF (ORTOLANG-TCOF), NCCFr (Torreira et al., 2010);
- Broadcast speech: EPAC (Estève et al., 2010);
- Professional speech: Voxpopuli (Wang et al., 2021).

Compared to MLS and Voxpopuli, our dataset is more diverse, carefully sourced and contains detailed metadata (speech type, and speaker gender). Moreover, compared to these, it has a more realistic representation of speech turns in real life. Statistics are reported in Table 1.



Corpus$_{License}$	# Utterances	Duration	# Speakers	Mean Utt. Duration	Speech type
Small dataset – 1K					
MLS French$_{CCBY4.0}$ [43]	**263,055** 124,590 / 138,465 / –	**1,096:43** 520:13 / 576:29 / –	**178** 80 / 98 / –	**15 s** 15 s / 15 s / –	Read
Medium-clean dataset – 2.7K					
EPAC**$_{NC}$ [45]	**623,250** 465,859 / 157,391 / –	**1,626:02** 1,240:10 / 385:52 / –	**Unk** – / – / –	**9 s** – / – / –	Radio Broadcasts
2.7k dataset total	**886,305** 590,449 / 295,856 / -	**2,722:45** 1,760:23 / 962:21 / -	-	-	-
Medium dataset – 3K					
African Accented French$_{Apache2.0}$ [46]	**16,402** 373 / 102 / 15,927	**18:56** – / – / 18:56	**232** 48 / 36 / 148	**4 s** – / – / –	Read
Att-Hack$_{CCBYNCND}$ [47]	**36,339** 16,564 / 19,775 / –	**27:02** 12:07 / 14:54 / –	**20** 9 / 11 / –	**2.7 s** 2.6 s / 2.7 s / –	Acted Emotional
CaFE$_{CCNC}$ [48]	**936** 468 / 468 / –	**1:09** 0:32 / 0:36 / –	**12** 6 / 6 / –	**4.4 s** 4.2 s / 4.7 s / –	Acted Emotional
CFPP2000$_{CCBYNCSA}$* [49]	**9853** 166 / 1,184 / 8,503	**16:26** 0:14 / 1:56 / 14:16	**49** 2 / 4 / 43	**6 s** 5 s / 5 s / 6 s	Spontaneous
ESLO2$_{NC}$ [50]	**62,918** 30,440 / 32,147 / 331	**34:12** 17:06 / 16:57 / 0:09	**190** 68 / 120 / 2	**1.9 s** 2 s / 1.9 s / 1.7 s	Spontaneous
GEMEP$_{NC}$ [51]	**1,236** 616 / 620 / –	**0:50** 0:24 / 0:26 / –	**10** 5 / 5 / –	**2.5 s** 2.4 s / 2.5 s / –	Acted Emotional
MPF [52], [53]	**19,527** 5,326 / 4,649 / 9,552	**19:06** 5:26 / 4:36 / 9:03	**114** 36 / 29 / 49	**3.5 s** 3.7 s / 3.6 s / 3.4 s	Spontaneous
PORTMEDIA$_{NC}$ (French) [54]	**19,627** 9,294 / 10,333 / –	**38:59** 19:08 / 19:50 / –	**193** 84 / 109 / –	**7.1 s** 7.4 s / 6.9 s / –	Acted telephone dialogue
TCOF$_{CCBYNCSA}$ (Adults) [55]	**58,722** 10,377 / 14,763 / 33,582	**53:59** 9:33 / 12:39 / 31:46	**749** 119 / 162 / 468	**3.3 s** 3.3 s / 3.1 s / 3.4 s	Spontaneous
Medium dataset total	**1,111,865** 664,073 / 379,897 / 67,895	**2,933:24** 1,824:53 / 1,034:15 / 74:10	-	-	-
Large dataset – 7K					
MaSS [56]	**8,219** 8,219 / – / –	**19:40** 19:40 / – / –	**Unk** – / – / –	**8.6 s** 8.6 s / – / –	Read
NCCFr$_{NC}$ [57]	**29,421** 14,570 / 13,922 / 929	**26:35** 12:44 / 12:59 / 00:50	**46** 24 / 21 / 1	**3 s** 3 s / 3 s / 3 s	Spontaneous
Voxpopuli$_{CC0}$ [44] *Unlabeled*	**568,338** – / – / –	**4,532:17** – / – / 4,532:17	**Unk** – / – / –	**29 s** – / – / –	Professional speech
Voxpopuli$_{CC0}$ [44] *transcribed*	**76.281** – / – / –	**211:57** – / – / 211:57	**327** – / – / –	**10 s** – / – / –	Professional speech
Large dataset total* **	**1,814,242** 682,322 / 388,217 / 99,084	**7,739:22** 1,853:02 / 1,041:07 / 4,845:07	-	-	-
Extra Large dataset – 14K					
Audiocite.net$_{CC-BY}$ [58]	**817 295** 425 033 / 159 691 / 232 571	**6698:35** 3477:24 / 1309:49 / 1911:21	**130** 35 / 32 / 63	**29 s** 29 s / 29 s / 29 s	Read
Niger-Mali Audio Collection $_{CCBYNCND}$ [59] [60]	**38 332** 18 546 / 19 786 / –	**111:01** 52:15 / 58:46 / –	**357** 192 / 165 / –	**10 s** 10 s / 10 s / –	Radio broadcasts
Extra Large dataset total	**2 669 869** 1 125 901 / 567 694 / 331 655	**14 548:58** 5 382:41 / 2 409:42 / 6 756:28	-	-	-

*Composed of audio files not included in the CEFC corpus v2.1, 02/2021; **speakers are not uniquely identified.; ***Stats of CFPP2000, MPF and TCOF have changed a bit due to a change in data extraction; License: CC=Creative Commons; NC=non-commercial; BY= Attribution; SA= Share Alike; ND = No Derivative works; CC0 = No Rights Reserved

Table 1 *Statistics for the speech corpora used to train SSL models according to gender information (male / female / unknown). The small dataset is from MLS only. Every dataset is composed of the previous one + additional data; MPF, TCOF and CFPP2000 appear twice with different stats as data extraction changed; duration: hour(s):minute(s)*

- **Pre-processing for SSL training:** Recordings were segmented using time stamps from transcriptions. We retrieved, when available, speaker labels and gender information. Following Baevski et al., 2020, we removed utterances shorter than 1s, and longer than 30s. When possible, overlapping speech sentences were also removed. When necessary, audio segments were converted to mono PCM 16bits, 16kHz.
- **Small dataset (approximately 1k hours):** It is only composed of the MLS corpus for comparison with Wav2Vec2.0 Baevski et al., 2020 which uses only read English speech. It is also gender balanced.

- **Medium dataset (approximately 3k hours):** It includes 2,933 hours of speech, from which 1,115 hours is read speech, 1,626 hours broadcast speech, 123 hours spontaneous speech, 38 hours acted telephone dialogues, and 29 hours acted emotional speech. Regarding gender, we collected 1,824 hours of speech from male speakers, 1,034 hours from female speakers, and 74 hours from unknown gender.
- **Large dataset (approximately 7.7k hours):** It has 4 additional corpora: MaSS, NCCFr and Voxpopuli (unlabeled + transcribed). It includes 7,739 hours of speech, from which 1,135 hours is read speech, 1,626 hours broadcast speech, 165 hours spontaneous speech, 38 hours acted telephone dialogues, 29 hours acted emotional speech, and 4744 hours professional speech. Except for NCCFr, no info about gender is given in the added datasets.
- **Extra-large dataset (approximately 14k hours):** It has 2 additional corpora audiocite.net and Niger-Mali Audio Collection. Audiocite.net includes freely shareable audiobooks of more than 6 600 hours. The Niger-Mali Audio Collection is data web-crawled from Studio Kalangou and Studio Tamani websites, with the authorization of Fondation Hirondelle.

2.3 Training and Sharing SSL Models

The *LeBenchmark* provides seven Wav2Vec2.0 models pretrained on the gathered French data described above. Following Baevski et al., 2020, two different Wav2Vec2.0 architectures (*large* and *base*) are coupled with our small (1K), medium (3K), large (7K) and extra-large (14K) corpora to form our set of Wav2Vec2.0 models: W2V2-Fr-1K-base, W2V2-Fr-1K-large, W2V2-Fr-3K-base, W2V2-Fr-3K-large, W2V2-Fr-7K-base and W2V2-Fr-7K-large. For the extra-large corpora, we use three architectures (light, large and xlarge) to form the W2V2-Fr-14K-light, W2V2-Fr-14K-large and W2V2-Fr-14K-xlarge.

Hyperparameters and architectures for base and large are identical to the ones first introduced in Baevski et al., 2020. Most models except 14K-xlarge have been trained on nodes equipped with four 32GB Nivida Tesla V100 hence triggering multi-node training to reach the desired 32 or 64 GPU. 14K-xlarge was trained with 80GB Nvidia Tesla A100 nodes equipped with eight GPU each. Data read and write operations were made throughout a fast Nested File

System (NFS) without any streaming library. A detailed summary of the hyperparameters used to train our SSL models can be found in Table 2. In practice, training is stopped at a round number of updates once the loss observed on the development set of the MLS corpus reaches a stable point. Pre-trained Wav2Vec2.0 models are shared with the community via HuggingFace for further integration with well-known toolkits such as SpeechBrain, Fairseq or Kaldi.

Pre-existing Wav2Vec2.0 models obtained from Fairseq are also considered in downstream experiments. First, *XLSR-53-large* is used as a comparison to multilingual models. Then, *W2V2-En-base* and *W2V2-En-large* (LS960) are used to assess English representations from LibriSpeech. For the sake of conciseness, we remove the prefix W2V2- from all our results tables in the next section.

Model	**Training Data**	**Transformer Blocks**	**Model Dimension**	**Inner Dimension**	**Heads**	**Updates**
Fr-1K-base	1,096 h	12	768	3,072	8	200K
Fr-1K-large	1,096 h	24	1024	4,096	16	200K
Fr-3K-base	2,933 h	12	768	3,072	8	500K
Fr-3K-large	2,933 h	24	1024	4,096	16	500K
Fr-7K-base	7,739 h	12	768	3,072	8	500K
Fr-7K-large	7,739 h	24	1024	4,096	16	500K
Fr-14K-light	14,000 h	12	512	3,072	8	500K
Fr-14K-large	14,000 h	24	1024	4,096	16	1M
Fr-14K-xlarge	14,000 h	48	1280	5,120	16	1M

Table 2 *Hyperparameters of our pre-trained SSL models*

3. *LeBenchmark* Results

Results on speech recognition, speech translation and other downstream tasks

We benchmark SSL models on four different tasks: Automatic Speech Recognition (ASR), Speech Language Understanding (SLU), Automatic Speech Translation (AST), and Named Entity Recognition (NER). Since our goal is to evaluate the impact of SSL for the best baselines for each task addressed, we have a different architecture for each task, and it corresponds to the best baseline performance we could obtain using MFCC/MFB features. As a different architecture/approach is used for each task, we evaluate the different SSL models as feature extractors for these tasks. These 'SSL extractors' are either 'task agnostic' or 'task specific' (SSL models fine-tuned on the task data), as further explained below.

3.1 Automatic Speech Recognition (ASR) Results

Automatic Speech Recognition (ASR) consists of transcribing the content of a speech utterance. In this section, we present ASR results using an end-to-end model and two datasets. Results focus on larger Wav2vec2.0 models (3K and 7K), as these are the ones for which we notice the most significant improvements.

- **Datasets:** The ASR tasks target two different types of corpora: Common Voice (Ardila et al. 2020) and ETAPE (Gravier et al. 2012). Common Voice is a very large crowd-sourced corpus (477 hours) of read speech in French with transcripts (train: 428h, dev: 24h, and test: 25h), while ETAPE is a smaller (36 hours) but more challenging corpus composed of diverse French TV broadcast programs (train: 22h, dev: 7h, and test: 7h).

- **Architecture:** Our models are implemented with the SpeechBrain toolkit (Ravanelli et al., 2021). The baseline system is fed by 80-dimension log Mel filterbank (MFB) features and is based on an encoder/decoder architecture with attention. When used with an SSL pre-trained Wav2Vec2.0 model, the system simply adds an additional hidden layer and an output layer on top of a Wav2Vec2.0 architecture.

- **Results:** Table 3 presents the results achieved with ASR systems on French Common Voice 6.1 and on ETAPE. Before the use of Wav2vec2.0 models for ASR, the baseline MFB-based system (first line of the table) was the state-of-the-art e2e model on CommonVoice/French. Other lines of the table present different Wav2vec2.0 models fine-tuned on labeled ASR data from CommonVoice or ETAPE. Wav2vec2.0 *base* and *large* models provided by ***LeBenchmark*** outperform clearly *En-large* and *XLSR-53-large* models. The best model is *Fr-3K-large*, pretrained on a smaller training dataset than *Fr-7K-large*, and it provides the best results on all the experiments.

Corpus	**CommonVoice**		**ETAPE**	
Features	**Dev**	**Test**	**Dev**	**Test**
MFB	17.67 (0.37)	20.59 (0.41)	54.03 (1.33)	54.36 (1.32)
En-large	12.05 (0.23)	14.17 (0.52)	42.14 (0.72)	44.82 (0.74)
XLSR-53-large	16.41 (0.27)	19.40 (0.29)	58.55 (0.65)	61.03 (0.70)
FR-1K-large	9.49 (0.20)	11.21 (0.23)	28.57 (0.79)	30.58 (0.88)
Fr-3K-base	11.25 (0.23)	13.22 (0.24)	26.14 (0.70)	28.86 (0.79)
Fr-3K-large	**8.00** (0.19)	**9.27** (0.20)	22.26 (0.76)	24.21 (0.85)
Fr-7K-base	10.84 (0.21)	12.88 (0.24)	25.13 (0.68)	28.16 (0.79)
Fr-7K-large	8.02 (0.18)	9.39 (0.21)	**21.34** (0.74)	**23.46** (0.83)
Fr-14K-light	19.86 (0.28)	22.81 (0.34)	58.30 (0.66)	59.82 (0.7)
Fr-14K-large	8.39 (0.19)	9.83 (0.21)	23.67 (0.81)	26.03 (0.89)
Fr-14K-xlarge	8.26 (0.19)	9.83 (0.21)	22.38 (0.95)	24.67 (0.83)

Table 3 *ASR results (WER%) on Common Voice and ETAPE corpora, with pre-trained Wav2vec2.0 models further fine-tuned on labeled ASR data. Gray numbers indicate 95% confidence intervals computed using bootstrap re-sampling as proposed in Bisani and Ney, 2004*

3.2 Automatic Speech Translation (AST) Results

Automatic speech-to-text translation (AST) consists of translating a speech utterance in a source language to a text in a target language. In this work, we are interested in translating directly from French speech to text in another language.

- **Dataset:** We selected subsets having French as the source in the multilingual TEDx dataset (Salesky et al., 2021). Our benchmark covers translation directions from French to three target languages: English (*en*), Spanish (*es*), and Portuguese (*pt*), with the following training sizes: 50h (*en*), 38h (*es*), and 25h (*pt*).
- **Experiments:** Our baselines are models using 80-dimensional MFB features. For learned representations derived from SSL models, we focused on the feature extraction approach where features are extracted from either task-agnostic or task-specific pre-training. Task-agnostic pre-training refers to the direct use of SSL models as feature extractors whereas the task-specific method consists of one additional phase where the SSL models are further trained on the in-domain task data, with (supervised fine-tuned) or without (self-supervised fine-tuned) labels.

  We performed supervised fine-tuning with speech transcriptions as labels and leave supervised fine-tuning with AST data for future work. In the task-specific scenario, we only considered three SSL models: two best French SSL models (*Fr-3K-large* and *Fr-7K-large*) and one best non-French SSL model (*XLSR-53-large*). Since the French speech is overlapped between the language pairs, we selected the pair having the most speech data (fr-en) to perform task-specific pre-training and used the obtained models to extract features for the remaining pairs (fr-es and fr-pt). For a fair comparison, we did not use additional data augmentation techniques or ASR encoder pre-training in the experiments.
- **Architecture:** We used a small Transformer (Vaswani et al., 2017) architecture having 6 layers of encoders, 3 layers of decoders, and hidden dimension 256 in all experiments. Following previous work (Nguyen et al. 2020; Evain et al. 2021-A), we inserted a block of Linear-ReLU before convolutional layers in the speech encoder for parameter efficiency and model performance reasons.

- **Results:** Table 4 displays the results of the AST experiments. One can observe that SSL features, whether task-agnostic or task-specific and whether being pre-trained on English, French, or multilingual data, outperform the baselines using MFB features by a large margin (except for the task-agnostic multilingual model XLSR-53 on the two pairs fr-es and fr-pt, which are in very low-resource settings).

  Comparing blocks: Among the three groups using SSL features (task-agnostic pre-training, task-specific self-supervised, and task-specific fine-tuned for ASR), the ASR fine-tuning approach (c) yields the best results. We observe considerable improvements from task-specific self-supervised (b) to task-specific fine-tuned (c) (+6.19, +8.50, +8.53 on average for en, es, and pt, respectively) while the benefits of using self-supervised fine-tuning compared to task-agnostic pre-training are only marginal or even slightly negative.

  The substantial gains when using the supervised fine-tuning approach (even with the somehow indirect signal of transcripts for the AST downstream task) shows that giving more signals of the task-specific data to the SSL models is helpful. In the case of task-specific self-supervised fine-tuning (b), we further trained the SSL models for 20k more steps on the raw task-specific data, whereas in ASR fine-tuned scenario (c), we used raw data plus the transcripts to guide the SSL models.

  Task-agnostic SSL: Focusing on task-agnostic block (a), we see that French SSL models clearly outperform those pre-trained on English and multilingual data. Multilingual XLSR-53 model surpasses the English models on fr-en, yet all of them fail to generate meaningful translations on fr-es and fr-pt where little training data is available.

  Comparing across different French SSL model sizes (base vs large), the large architecture yields considerable improvements (nearly 3 to 6 BLEU points) over its base counterpart. When looking into the French SSL models with different amounts of pre-training data (1K, 3K, and 7K), we observe large gains for the base architecture from using 1K to using 3K or more pre-training data. There is, however, no significant difference between base models using 3K and 7K data. Using 7K data even hurts the performance on the pair fr-pt. On the other hand, for the large network, using more data consistently improves the performance on all language pairs.

Task-specific SSL: Finally, moving on to task-specific models, Fr-7K-large is the best-performing model (or being on par with the best one) in each group. Noticeably, there is a huge improvement when using the ASR fine-tuning approach (c) for the multilingual XLSR-53 model. The method considerably boosts the performance of the multilingual model (compared to using it directly or further pre-training it on the task data) and makes it even on par with the best French SSL models.

Features	**Valid**			**Test**		
	en	**es**	**pt**	**en**	**es**	**pt**
MFB	1.15 (0.27)	0.67 (0.15)	0.61 (0.13)	1.10 (0.14)	0.87 (0.12)	0.32 (0.03)
(a) Task agnostic pre-training						
En-base	5.54 (0.27)	1.30 (0.17)	0.54 (0.11)	5.20 (0.28)	1.47 (0.15)	0.38 (0.05)
En-large	4.11 (0.25)	1.67 (0.20)	0.32 (0.03)	3.56 (0.22)	2.29 (0.18)	0.43 (0.05)
Fr-3K-base	15.05 (0.49)	13.19 (0.25)	4.44 (0.29)	14.80 (0.47)	14.27 (0.44)	4.72 (0.25)
Fr-3K-large	17.94 (0.51)	16.40 (0.49)	8.64 (0.34)	18.00 (0.51)	18.12 (0.48)	9.55 (0.36)
Fr-7K-base	15.13 (0.45)	12.78 (0.40)	2.65 (0.20)	14.50 (0.45)	13.61 (0.44)	2.66 (0.23)
Fr-7K-large	19.23 (0.54)	17.59 (0.49)	9.68 (0.37)	19.04 (0.53)	18.24 (0.49)	10.98 (0.41)
Fr-14K-light	10.31 (0.38)	9.83 (0.33)	4.96 (0.31)	10.92 (0.43)	10.52 (0.42)	5.79 (0.33)
Fr-14K-large	18.93 (0.40)	17.22 (0.41)	9.03 (0.35)	18.97 (0.47)	18.12 (0.42)	10.11 (0.39)
Fr-14K-xlarge	18.14 (0.42)	15.90 (0.39)	5.46 (0.29)	18.35 (0.48)	17.19 (0.43)	6.59 (0.35)
XLSR-53-large	7.81 (0.33)	0.49 (0.13)	0.43 (0.07)	6.75 (0.29)	0.52 (0.08)	0.36 (0.05)
(b) Task specific pre-training (self-supervised on mTEDx)						
Fr-3K-large	18.54 (0.53)	16.40 (0.48)	8.81 (0.36)	18.38 (0.52)	17.84 (0.48)	10.57 (0.41)
Fr-7K-large	19.65 (0.55)	17.53 (0.47)	9.35 (0.36)	19.36 (0.54)	18.95 (0.53)	10.94 (0.38)
Fr-14K-light	6.5 (0.27)	5.7 (0.27)	3.0 (0.21)	5.9 (0.28)	5.7 (0.26)	2.9 (0.17)

Fr-14K-large	23.6 (0.59)	23.3 (0.58)	18.7 (0.54)	23.1 (0.55)	24.2 (0.62)	21.8 (0.58)
Fr-14K-xlarge	25.1 (0.59)	23.7 (0.56)	20.7 (0.58)	24.4 (0.60)	25.5 (0.59)	23.7 (0.62)
XLSR-53-large	15.6 (0.49)	15.6 (0.45)	8.4 (0.31)	12.5 (0.47)	15.8 (0.44)	9.1 (0.36)
(c) Task specific pre-training (fine-tuned for ASR on mTEDx)						
Fr-3K-large	21.09 (0.53)	19.28 (0.53)	14.40 (0.47)	21.34 (0.58)	21.18 (0.52)	16.66 (0.49)
Fr-7K-large	**21.41** (0.51)	20.32 (0.49)	**15.14** (0.48)	**21.69** (0.58)	**21.57** (0.52)	**17.43** (0.52)
XLSR-53-large	21.09 (0.54)	**20.38** (0.56)	14.56 (0.45)	20.68 (0.53)	21.14 (0.55)	17.21 (0.54)

Table 4 *BLEU on valid and test sets of multilingual TEDx (mTEDx). The highest value in each group (task-agnostic pre-training, task-specific self-supervised, and supervised fine-tuning) is underlined while the best value in each column is highlighted in bold. Gray numbers denote the standard deviation computed using bootstrap re-sampling (Koehn et al. 2004)*

3.3 Spoken Language Understanding (SLU) Results

Spoken Language Understanding (SLU) aims at extracting a semantic representation from a speech signal in human-computer interaction applications (De Mori, 1997). Given the difficulty of creating an open-domain SLU application, many works focus on specific domains. We focus on the hotel information and reservation domain provided within the French corpus MEDIA (Bonneau Maylard et al., 2006; Quarteroni et al., 2009).

- **Dataset:** The MEDIA corpus is made of 1~250 human-machine dialogues acquired with a *Wizard-of-Oz* approach, where 250 users followed 5 different reservation scenarios. Spoken data were manually transcribed and annotated with domain concepts, following a rich ontology. The official corpus split is made up of 12,908 utterances (41.5 hours) for training, 1,259 utterances (3.5 hours) for development and 3,005 utterances (11.3 hours) for test. We note that while all turns have been manually transcribed and can be used to train ASR models, only user turns have been annotated with concepts and can be used to train SLU models. This results in only 41.5 hours of speech training data for ASR models and only 16.8 hours for SLU models.
- **Architecture:** All our models are based on LSTM (Hochreiter and Schmidhuber, 1997) seq2seq with attention (Bahdanau et al., 2014), being like the one proposed in previous

works (Dinarelli et al., 2017; Dinarelli et al., 2020, Evain et al., 2021-A). We use a similar speech encoder employing a pyramidal hierarchy of RNN layers like Chan et al., 2016 and Evains et al., 2021.

The decoder has been also improved, integrating two attention mechanisms: one as usual for attending the encoder's hidden states; the other for attending all previous decoder prediction's embeddings, instead of the previous prediction only like in the original LSTM-based encoder-decoder models (Bahdanau et al., 2014). Our model is implemented using the *Fairseq* library (Ott et al., 2019).

- **Experiments:** We use a total of 3 bidirectional LSTM layers of size 256 stacked in a pyramidal fashion in our encoder and the LSTM decoder has 2 layers of size 256. In addition to using spectrogram features and features from task agnostic SSL models, we also use features from task specific models (SLU on MEDIA). Two types of task-specific pre-training are performed: *self-supervised* which consists in resuming the SSL model training using the MEDIA training data and minimizing the Wav2Vec 2.0 loss ((b) self-supervised on MEDIA in the results table, also called task-adaptive pre-training in Gururangan et al., 2020); and *ASR supervised* ((c) fine-tuned for ASR on MEDIA in the results table) which consists in fine-tuning the full SSL model for a supervised downstream task with a CTC loss minimization objective (Graves et al., 2006).

  Finally, in this work we chose to fine-tune models with respect to the ASR task on MEDIA (not the SLU one) to see how it compares to self-supervised fine-tuning. We leave fine-tuning with respect to SLU for future work.

- **Results:** The results for SLU obtained with different speech representations are shown in Table 5. They are given in terms of Concept Error Rate (CER), computed the same way as Word Error Rate (WER) but on concept sequences. CER are accompanied by standard deviations (in gray), computed with the bootstrap method of Bisani and Ney, 2004 .

  We first note that our *spectrogram* baseline obtains a substantial improvement over the one in Evain et al., 2021-A . Such gain is due to an optimization of settings and the model architecture. Using SSL model features as input resulted in a significant reduction in CER, even when using English SSL models (CER from 31.10 to 20.84 on the test set with the *base* model).

Task-agnostic SSL: In the best scenario among task-agnostic pre-trained models, we achieved a CER of 15.95 on the test data with *Fr-3K-large* features. Surprisingly, using features from the model trained with 7k hours of speech (*Fr-7K-large*), results are worse on both dev and test. In contrast, we also evaluated these models in terms of ASR performance, finding that the 7k-model led to the best results.

Task-specific SSL: We performed task-specific pre-training only with the most effective SSL models: French 3k and 7k models and multi-lingual *XLSR-53-large*. The best overall pre-trained model is the 7k-model fine-tuned for ASR on MEDIA, though results are close to those obtained with features from the 3k-model (13.97 vs. 13.78). Indeed, our significance tests confirm that these two models are equivalent, and they are significantly better than all the others. This shows that pre-trained SSL speech models can be specialized using task specific pre-training with either self-supervised learning on raw speech (block (b) in the table) or fine-tuning on raw speech and associated transcripts (block (c) in the table), the latter being slightly better than the former.

Features	**Dev**	**Test**
Spectrogram from Evain et al., 2021-A	33.63 (1.28)	34.76 (0.83)
spectrogram	**29.07** (1.31)	**31.10** (0.83)
(a) Task agnostic pre-training		
En-base	22.38 (1.24)	20.84 (0.68)
En-large	23.31 (1.31)	25.26 (0.77)
Fr-1K-base	22.89 (1.26)	23.27 (0.76)
Fr-1K-large	20.10 (1.10)	20.66 (0.72)
Fr-3K-base	19.44 (1.11)	18.56 (0.67)
Fr-3K-large	**15.96** (1.02)	**15.95** (0.62)
Fr-7K-base	20.70 (1.07)	18.86 (0.68)
Fr-7K-large	17.25 (1.02)	16.35 (0.66)
XLSR-53-large	18.45 (1.15)	18.78 (0.66)
(b) Task specific pre-training (self-supervised on MEDIA)		
Fr-3K-large	15.93 (1.01)	**14.94** (0.60)
Fr-7K-large	**15.42** (1.03)	15.17 (0.60)
XLSR-53-large	16.77 (1.09)	15.56 (0.61)
(c) Task specific pre-training (fine-tuned for ASR on MEDIA)		
Fr-3K-large	**14.49** (1.06)	13.97 (0.59)
Fr-7K-large	14.58 (1.01)	**13.78** (0.58)
XLSR-53-large	16.05 (1.05)	15.46 (0.60)

Table 5 *End-to-end SLU decoding results (Concept Error Rate %) on the MEDIA corpus*

3.4 Named Entity Recognition (NER) Results

Named Entity Recognition (NER) aims to locate and classify named entity mentions in speech transcripts into pre-defined categories (such as person names, organizations, locations, …).

- **Dataset:** The QUAERO data has been developed during the research project QUAERO (2008-2013). It consists of the manual annotation of named entities of the manual transcription of the ESTER1 corpus. ESTER1 Graves et al., 2004 is an evaluation campaign focusing on the evaluation of orthographic transcription, event detection and tracking, and information extraction. An official QUAERO test dataset has also been added. This entire corpus is composed of data recorded from French radio and TV stations between 1998 and 2004. The official corpus split is made up of 93.5 hours for training and 6.5 hours for testing. Named Entities often include seven major groups: person, location, organization, amount, time, production and function. Within the framework of the QUAERO project, an extended named entity annotation with compositional and hierarchical structure has been proposed (Galibert et al., 2011). The QUAERO dataset does not contain a development dataset. So, we use the ETAPE development part. ETAPE is a French dataset composed of data recorded from French radio and TV stations between 2010 and 2011. It is annotated with the same pre-defined categories of entities used in the QUAERO annotation.
- **Architecture:** Our model is based on end-to-end approaches. The end-to-end system is composed of a large pre-trained French wav2vec model (LeBenchmark Fr-7K-large), a linear hidden layer of 1024 units, and a softmax output layer. The loss function used for the supervised fine-tuning step is the Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006).
- **Results:** The obtained results for NER are shown in Table 6. They are given in terms of Entity Error Rate (EnER), computed in the same way as Word Error Rate (WER) but only on entity sequences, exactly like the Concept Error Rate used for SLU. We compute the total EnER '*All Entities*' and an EnER for each entity category. We report also results in terms of WER for the transcription without entities. The results are obtained by using the flat version of the named entity representation retained in the QUAERO dataset (i.e., not using a structured representation).

Features	Test
(a) Word Error Rate (WER)	
WER	10.9%
(b) Entity Error Rate (EnER)	
All Entities	32.24%
Entity 'Person'	27.60%
Entity 'function'	52.84%
Entity 'organisation'	46.24%
Entity 'location'	27.09%
Entity 'production'	70.87%
Entity 'amount'	24.66%
Entity 'time'	28.8%

Table 6 *End-to-end NER decoding results (Entity Error Rate %) on the QUAERO dataset*

4. SELMA-19: the SELMA multilingual wav2vec 2.0 model

Thanks to Deutsche Welle, the SELMA project could access a large amount of multilingual speech data, specifically audio or video documents related to news. Thanks to our experience coming from the LeBenchmark initiative and from the literature (Hsu et al., 2021) we expected that a wav2vec 2.0 pretrain on in-domain data (here, journalistic data) achieves a better performance on this kind of data. So, we trained a multilingual model using journalistic data. We present in the next subsections the data collection and preparation and the results.

4.1 Data collection and preparation

To pretrain by self-supervision a wav2vec, we collected a huge amount of multilingual data provided by the Deutsche Welle partner. Statistics of the raw data are presented in the table below:

Language	Type	Items	Duration	Date range
Arabic	Audio	151	108 h	2021-08-06 > 2022-07-31
	Video	6,15	965 h	2021-08-06 > 2022-07-31
	All	**6,301**	**1,073 h**	
Brazilian	Audio	963	90 h	2018-07-24 > 2022-07-30
	Video	1,957	209 h	2018-07-24 > 2022-07-30
	All	**2,92**	**299 h**	
Chinese	Audio	640	279 h	2012-11-18 > 2022-07-31
	Video	2,922	129 h	2012-11-18 > 2022-07-31
	All	**3,562**	**408 h**	
Dari	Audio	455	48 h	2017-08-15 > 2022-12-06
	Video	721	85 h	2017-08-15 > 2022-12-06
	All	**1,176**	**133 h**	
English	Article	903	2 h	2021-01-02 > 2022-07-31
	Audio	1,063	433 h	2021-01-02 > 2022-07-31
	All	**1,966**	**435 h**	
French	Audio	1,852	661 h	2020-06-05 > 2022-07-31
	Video	974	92 h	2020-06-05 > 2022-07-31
	All	**2,826**	**752 h**	

German	Audio	99	34 h	2022-04-01 > 2022-07-31
	Video	1,463	244 h	2022-04-01 > 2022-07-31
	All	**1,562**	**279 h**	
Greek	Audio	1,234	52 h	2013-06-05 > 2022-12-11
	Video	1,333	69 h	2013-06-05 > 2022-12-11
	All	**2,567**	**121 h**	
Hausa	Audio	10,763	6,252 h	2013-12-10 > 2022-07-31
	Video	675	32 h	2013-12-10 > 2022-07-31
	All	**11,438**	**6,284 h**	
Hindi	Audio	50	6 h	2013-01-06 > 2022-07-31
	Video	3,985	297 h	2013-01-06 > 2022-07-31
	All	**4,035**	**303 h**	
Indonesian	Video	3,019	232 h	2013-06-10 > 2022-07-31
	All	**3,019**	**232 h**	
Pashto	Audio	1,06	89 h	2013-06-28 > 2022-12-13
	Video	847	89 h	2013-06-28 > 2022-12-13
	All	**1,907**	**177 h**	
Persian	Audio	1,824	375 h	2012-02-27 > 2022-07-31
	Video	2,41	112 h	2012-02-27 > 2022-07-31
	All	**4,234**	**487 h**	
Polish	Audio	46	8 h	2012-12-20 > 2022-12-13
	Video	2,414	139 h	2012-12-20 > 2022-12-13
	All	**2,46**	**148 h**	
Russian	Audio	98	33 h	2011-07-26 > 2022-07-31
	Video	12,98	1,281 h	2011-07-26 > 2022-07-31
	All	**13,078**	**1,314 h**	
Spanish	Audio	70	110 h	2021-01-01 > 2022-07-31
	Video	7,109	951 h	2021-01-01 > 2022-07-31
	All	**7,179**	**1,061 h**	
Turkish	Audio	5,257	713 h	2011-08-05 > 2022-07-31
	Video	7,485	672 h	2011-08-05 > 2022-07-31
	All	**12,742**	**1,385 h**	
Ukrainian	Video	9,233	551 h	2012-12-05 > 2022-07-29
	All	**9,233**	**551 h**	
Urdu	Audio	2,769	302 h	2012-05-29 > 2022-10-14
	All	**2,769**	**302 h**	
All	**All**	**94,974**	**15,743 h**	

***Table** 7 Statistics of raw data shared by Deutsche Welle to be used to pretrained a multilingual wav2vec 2.0 model*

These files have been processed to extract only speech segments and to specify the gender of the speaker involved for each speech segment. We **aim to build a gender-balanced and language-balanced pretraining data: we kept a maximum of 250 hours of speech for each language (for a total of about 3,000 hours of speech).**

The SpeechBrain toolkit, interfaced to the HuggingFace *transformers* library, was used for this SSL training.

4.2 First results

We trained two models: one from scratch, the other one by continuing the training of the Meta wav2vec2 XLS-R-128 (large architecture) on our gender-balanced and language-balanced data. We observed that even if the model pretrained from scratch on our 3,000 hours reached interesting results for automatic speech recognition, it is not competitive in comparison to the XLS-R-128 model pretrained on more than 60,000 hours. Our second model (SELMA-19-XLSR-128) that continues the training of XLS-R-128, got slightly better results than the original one, but these improvements were not significant. However, it is interesting to note that fine-tuning the SSL model on gender-balanced data has an impact on the performance achieved with female voices, as illustrated in the table below:

Model =	SELMA-19-XLSR-128				XLS-R-128		
Train set =	female only	male only	mixed genders		female only	male only	mixed genders
Test WER	**19,4**	19,2	**18,5**		19,7	19,2	18,7
Test WER on female subset	22,3	24,0	**21,8**		**21,8**	**23,9**	22,2
Test WER on male subset	**20,9**	20,2	**19,3**		21,2	20,2	20,0
Delta F-M	1,4	3,8	2,5		0,6	3,7	2,2

Table 8 *Results using the SELMA-19-XLSR-128 model*

These experiments were carried out on the CommonVoice10 corpus on French speech.

Our models are released under a free and very permissive licence to contribute to the advances of the research community. It is accessible on the SELMA account on the HuggingFace platform: https://huggingface.co/H2020SELMA

5. Speech Synthesis

Text to speech (TTS), or speech synthesis, which aims to synthesize intelligible and natural speech given text, is a hot research topic in speech, language, and machine learning communities. Thanks to the advances in deep learning and artificial intelligence, neural network-based TTS has significantly improved the quality of synthesized speech in recent years.

In this section, the neural network-based architecture developed in SELMA for our first text-to-speech engine is presented, in addition to the data used for the training process. Lastly, we discuss how our work on speech synthesis, applied to Brazilian Portuguese broadcast news, could be evaluated.

5.1 Architecture

Our TTS system consists of two components, an acoustic model and a vocoder. The acoustic model generates acoustic features from linguistic features (in this case: text), and the vocoder synthesizes waveform from the acoustic features.

For the acoustic model, we conducted experiments with several architectures. This allowed us to draw the following conclusion: purely in terms of quality and naturalness Tacotron 2 [Shen et al, 2018] + DDC gave us the best performance. Other architectures like GlowTTS [Kim et al., 2020], SpeedySpeech [Vainer and Dusek, 2020] or FastSpeech [Ren et al., 2019] are faster and synthesize intelligible speech but not as good as Tacotron 2.

Considering the vocoder, we also had multiple choices and primarily focused on two architectures: Hifi-Gan [Kong et al., 2020] and WaveRNN [Kalchbrenner et al., 2018]. The first one did not give us the expected results, so we have decided to go for the second one. From the paper, there is not a significant difference between the two in terms of speech quality, the main difference is about inference time. However, since we have no inference real-time constraints, this is not a problem.

5.2 Data

We use the audio news bulletins that are produced by DW's Brazil department to train the speech synthesis engine. The audio files have been downloaded from YouTube and the scripts were retrieved from GitHub in a repository with all the text scripts that DW uses to produce their weekday news podcasts.

The dataset contains approximately 32 hours of speech from 8 speakers. The repartition of utterances and hours per speaker after cleaning is described below in Table 7.

#	Name	Training utterances	Hours
1	Roberto	3510	8.5
2	Alexandre	3348	7.7
3	Philip	2759	6.0
4	Leila	2077	5.1
5	Bruno	679	1.7
6	Marcio	554	1.3
7	Clarissa	357	0.9
8	Renate	295	0.7

Table 9 *Repartition of utterances and hours per speaker*

Additional speech synthesis models within the SELMA project have been trained also for Urdu and Latvian languages following the same recipe developed on the Brazilian Portuguese data described here. The Urdu speech synthesis was trained on DW data, and the Latvian speech synthesis was trained on the Latvian blind people audiobooks library data.

5.3 Evaluation

Currently, we are still working on the evaluation part of the speech synthesis engine. The evaluation protocol can be divided into two parts. First, we will evaluate the accuracy of the speech synthesis using a speech recognition model.

Using the original transcription and the output of an ASR model, we can compute the Word Error Rate (WER) which is a common metric for measuring speech-to-text accuracy of automatic speech recognition systems.

As this first evaluation protocol doesn't measure the prosodic aspect of the TTS system, we introduced a second one involving human rating. The next step will be to organize a perceptual evaluation campaign where samples are rated by humans on a scale from 1 to 5 with 0.5-point increments, from which a subjective mean opinion score (MOS) is calculated.

A Mean Opinion Score (MOS) is a numerical measure of the human-judged overall quality of an event or experience. In telecommunications, a Mean Opinion Score is a ranking of the quality of voice and video sessions.

A demo webpage with our first system here: click here to access our TTS demonstration webpage

6. Hybrid ASR system

Hybrid automatic speech recognition systems are based on HMM/DNN acoustic models of phonemes, a dictionary of words with their explicit pronunciations (sequence of phonemes), and language models.

Kaldi is a popular open-source toolkit designed to build such ASR systems. In SELMA, we implemented ASR systems for different languages using Kaldi, mainly to be integrated into the SELMA platform as the first ASR components.

6.1 French ASR

A Kaldi-based ASR system has been built for the French language. The acoustic models (AM) are trained on 40-dimensional high-resolution (hires) MFCC features with a state-of-the-art factorized time delay neural network (TDNN-F) architecture (Povey et al., 2018 ; Peddinti et al., 2015) on 300 hours of French Broadcast data with manual transcriptions. The acoustic model was trained using lattice-free maximum mutual information (LF-MMI) (Povey et al., 2016) and cross-entropy criteria. Speed and volume perturbation have been applied for data augmentation (Ko et al., 2015). The word error rate got on Broadcast News data not included in the training data is around 17.5%.

6.2 Latvian ASR

The baseline ASR system for Latvian is trained using the Kaldi framework. The acoustic model has been trained on a general-domain Latvian speech corpus containing 100 hours of broadcast recordings (Pinnis at al., 2014) augmented with various noisy recordings and musical recordings from the MUSAN corpus (Snyder, 2015). The TDNN+LSTM neural network is trained on 40-dimension FBANK vectors. Language models (LM) are trained using the SRILM toolkit (Stolcke, 2002). Trigram language models pruned to 1e-8 are used in all experiments. The LM is trained on the Latvian portion of the CommonCrawl. A rule-based system is used to generate the pronunciation lexicon based on 52 phonemes. The word error rate (WER) is measured on 22 minutes of various radio and TV broadcasts and is around 10.5%.

6.3 English, German, Spanish, Arabic ASR

Kaldi-based ASR systems for English, German, Spanish and Arabic have been developed by various partners (University of Edinburgh, IDIAP, QCRI) in the H2020 SUMMA project (Grant agreement: 688139) and released publicly afterwards.

These legacy systems have been adapted for use in the SELMA project as baseline ASR systems, although technical incompatibility with the latest Kaldi versions and high WERs of around 20% on broadcast news limited the scope of their use.

7. Spoken language understanding in low resource scenarios

In our low resource SLU scenario, an end-to-end model for ASR and a corpus of textual documents with named entity annotations but without the corresponding audios are available.

7.1 End-to-end model for named entity recognition from speech without paired training data

Our approach (Mdhaffar et al., 2022) is based on the use of an external model trained to generate a sequence of vectorial representations from text. These representations mimic the hidden representations that could be generated inside an end-to-end automatic speech recognition model by processing a speech signal. A SLU neural module is then trained to use these representations as input and the annotated text as output. Last, the SLU module replaces the top layers of the ASR model to achieve the construction of the end-to-end model.

To generate the simulated ASR hidden representations (or ASR embeddings), we train a sequence-to-sequence neural model, called *Text-to-ASR-Embeddings* model. Such an approach can be compared to propositions in literature that use synthetic voices to feed an ASR end-to-end model.

We motivated our proposition for different reasons. First, the use of synthetic speech introduces some artifacts in the input of the ASR model. If the ASR model is fine-tuned on such synthetic voices, these artifacts will degrade the capability of the model to process natural voices. A solution to avoid this consists of freezing the weights of the bottom layers and only update the weights of the higher layers, in which the semantic is better encoded. Since the bottom layers were optimized to process natural speech, the quality of the embeddings computed from synthetic speech is not guaranteed and can introduce a gap between embedding computed from natural and computed from synthetic speech.

With our approach, we aim to reduce this gap. In addition, our approach needs less computation at training time than the ones based on synthetic speech, since we avoid the use of a consequent number of lower layers.

To train this *Text-to-ASR-Embeddings* neural model, we must produce a training dataset composed of pairs of transcriptions, used as input, and sequences of ASR embeddings, used as output. To produce this training dataset, the end-to-end ASR model is used to transcribe its training dataset. For each transcribed utterance, we extract a sequence of ASR embeddings from a hidden layer, and associate this ASR embedding sequence to the automatic transcription. When the entire ASR training data has been processed, the ASR embedding sequences and their associated automatic transcriptions are used to train the *Text-to-ASR-Embeddings* model, as illustrated in (A) in the following figure.

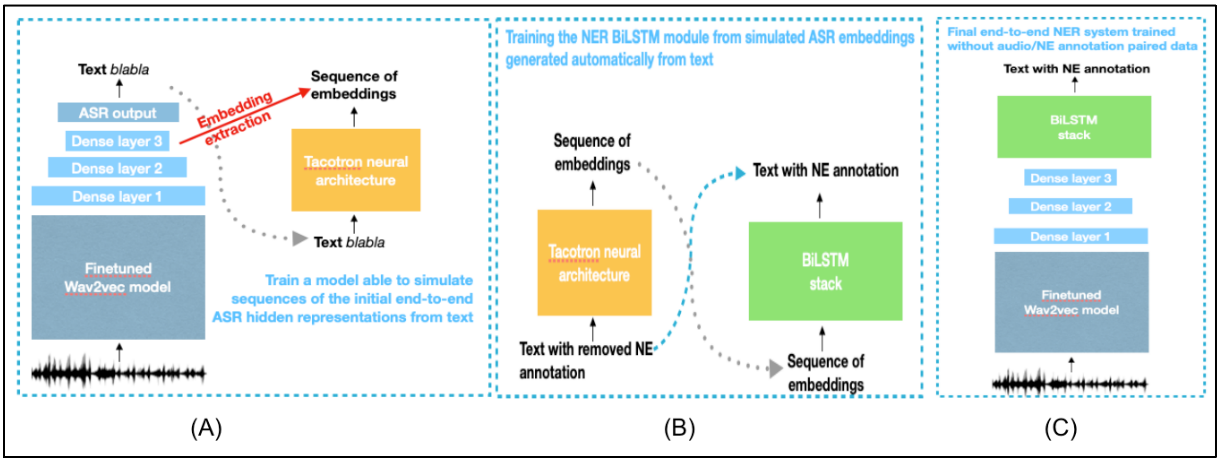

Text *blabla*
ASR output
Dense layer 3
Dense layer 2
Dense layer 1
Finetuned Wav2vec model
Embedding extraction
Sequence of embeddings
Tacotron neural architecture
Text *blabla*
Train a model able to simulate sequences of the initial end-to-end ASR hidden representations from text
(A)
Training the NER BiLSTM module from simulated ASR embeddings generated automatically from text
Sequence of embeddings
Tacotron neural architecture
Text with removed NE annotation
Text with NE annotation
BiLSTM stack
Sequence of embeddings
(B)
Final end-to-end NER system trained without audio/NE annotation paired data
Text with NE annotation
BiLSTM stack
Dense layer 3
Dense layer 2
Dense layer 1
Finetuned Wav2vec model
(C)

Figure 1 *End-to-End model for named entity recognition from speech without paired data*

At this stage, we obtain a module able to simulate ASR embeddings from text. Our objective is then to train a neural SLU sub-module able to convert such a sequence of ASR embedding into an automatic transcription with SLU annotation, like annotation of named entities.

For this purpose, we exploit the textual dataset with semantic annotation. For each sentence in this dataset, we first remove the semantic annotation to keep only the sequence of words. Thanks to the *Text-to-ASR-Embeddings* model, we transform this sequence of words to a sequence of ASR embeddings (B). We iterate this process for all the annotated sentences in the semantic textual dataset. We get a set of pairs composed of a sequence of ASR embeddings and the corresponding text sequence of words semantically annotated. Once the entire textual dataset has been processed, we use this data to train an SLU sub-module able to generate a sequence of words semantically annotated from a sequence of ASR embeddings (B).

Finally, we plug the end-to-end ASR and the SLU sub-module (C). In order to merge the ASR model with the SLU sub-module, we keep all the ASR hidden layers needed to generate the ASR embeddings that can be mimicked by the *Text-to-ASR-Embeddings* model. The mimicked hidden layer is then connected to the SLU sub-module.

The final model is an end-to-end model able to transcribe and extract semantic information from speech, while no real paired training data exists.

Our approach, based on artificial ASR embeddings generated from text, exhibits highly promising results outperforming alternative approaches based on the use of synthetic speech. These results, computed in terms of Name Entity Error Rate (NEER) are presented in the Table below.

Training data	**Dev**	**Test**
ASR embeddings simulation (ours)	47.6	39.1
Synthetic speech (all weights are updated)	65.2	62.7
Synthetic speech (frozen speech encoder)	86.4	92.5
Oracle (real audio)	45.9	34.1

Table 10 *Evaluation in NEER (%) of our approach to train an end-to-end NER model without paired training data compared to other approaches using speech synthesis, and compared to the ideal scenario when paired data is available*

We consider that this approach can be extended to similar SLU tasks such as slot filling and opens new perspectives in different use cases where enriching or adapting the linguistic knowledge captured by an end-to-end ASR model is needed.

7.2 Language portability of spoken language understanding model

SAMU-XLSR is based on the pre-trained multilingual XLSR on top of which all the embeddings generated by processing an audio file are connected to an attentive pooling module. Thanks to this pooling mechanism (which is followed by linear projection layer and the *tanh* function), the frame-level contextual representations are transformed into a single utterance-level embedding vector. Figure 1 summarizes the training process of the SAMU-XLSR model.

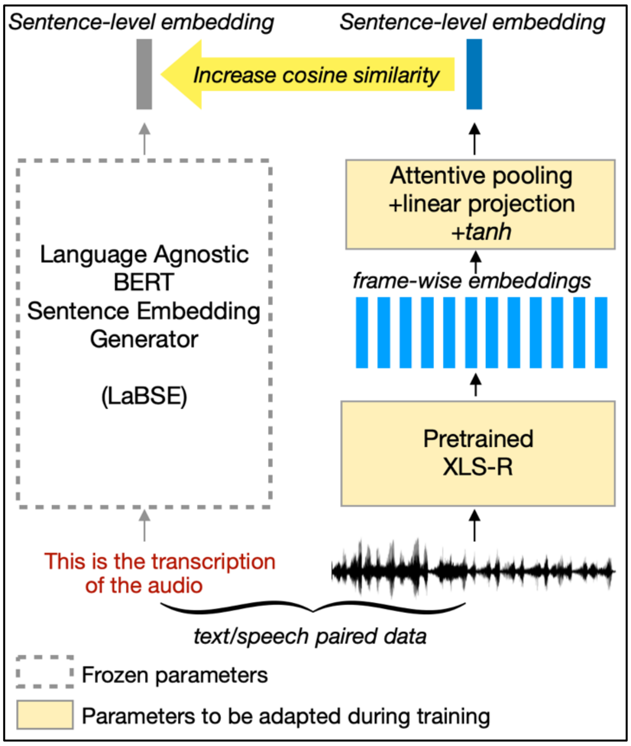

Sentence-level embedding
Sentence-level embedding
Increase cosine similarity
Language Agnostic BERT Sentence Embedding Generator
(LaBSE)
Attentive pooling +linear projection +tanh
frame-wise embeddings
Pretrained XLS-R
This is the transcription of the audio
text/speech paired data
Frozen parameters
Parameters to be adapted during training

***Figure 2** Training SAMU-XLSR*

Notice that the weights from the pre-trained XLS-R model continue being updated during the process. The utterance-level embedding vector of SAMU-XLSR is trained via knowledge distillation from the pre-trained language agnostic LaBSE model (Feng et al., 2022). The LaBSE model has been trained on 109 languages and its text embedding space is semantically aligned across these 109 languages. LaBSE attains state-of-the-art performance on various bi-text retrieval/mining tasks, while yielding promising zero-shot performance for languages not included in the training set (probably thanks to language similarities).

Thus, given a spoken utterance, the parameters of SAMU-XLSR are trained to accurately predict a text embedding provided by the LaBSE text encoder of its corresponding transcript.

During Y2, we investigate the use of SAMU-XLSR in order to train a Spoken Language Understanding neural model in French language and transfer it to the Italian language for which the amount of annotated data related to the SLU task (extraction of semantic concepts/values for hotel reservation task-oriented human/machine dialogue) is very low (less than 8 hours) or zero. In the zero-shot scenario (the model is trained on French and evaluated on Italian), our SAMU-XLSR-based model can get a Concept Error Rate (CER) of 54.6% while a classical XSLR model get 85.3%. When a few data in Italian is available, the gain in CER is less significant (26.2% instead of 26.9%), but the gain in Word Error Rate is promising (17.8% instead of 20%).

More details are available in (Laperrière et al., 2023).

7.3 Application to the Tunisian Dialect: creation of the TARIC-SLU corpus and speech encoder benchmark

Creation of the TARIC-SLU dataset

Our dataset for SLU is sourced from the TARIC dataset (Masmoudi et al, 2014). TARIC dataset was dedicated to training and evaluating Tunisian Dialect Automatic Speech Recognition in the context of human-to-human dialogues for train reservation task.

The acquisition of the TARIC dataset was carried out in some train stations in Tunisia.

Overall, TARIC dataset comprises 4,000 oral dialogue recordings from 108 speakers along with their manual transcriptions. In (Mdhaffar et al, 2024), TARIC dataset has been augmented with a semantic annotation. As detailed below, semantic annotation includes labelling with speech acts and slots/values. We annotated the TARIC-SLU dataset with two levels of labels: (1) speech act and (2) slot-value labels.

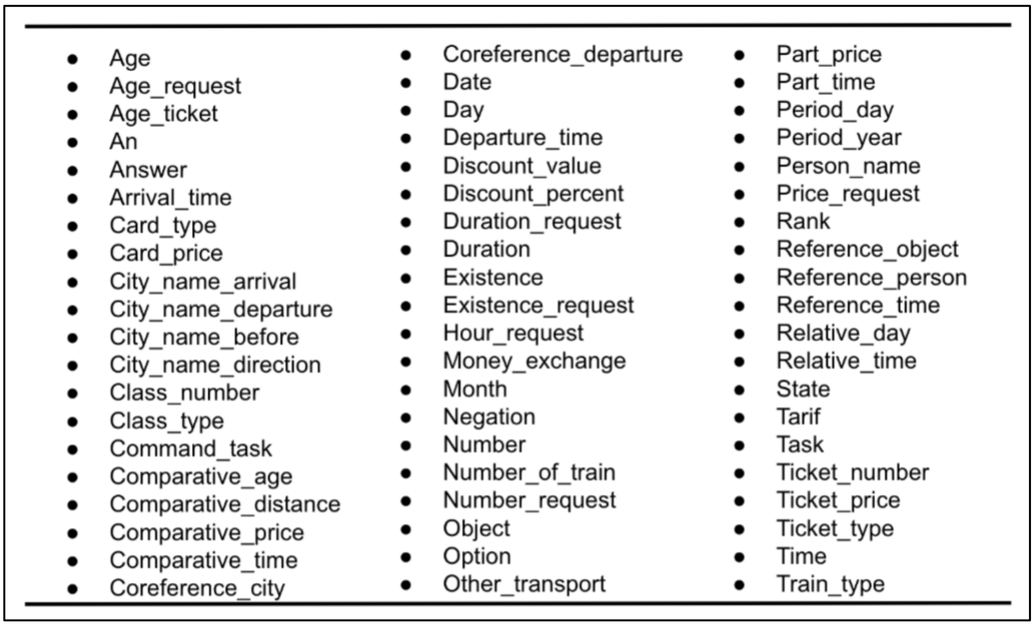
TARIC-SLU has been meticulously annotated with three distinct speech acts, signifying a comprehensive understanding of the main communicative intentions within the text.

- Directive query: is used when a user asks for information or makes a request.
- Directive answer: is the speech act where the speaker responds to a query or request for information by providing a specific answer or solution.

- Politeness: represents greetings at the beginning or during the conversation. It also includes apologizing, congratulating, thanking, commiserating, and expressing gratitude or good wishes.

The semantic representation employs a slot-value structure.

A semantic segment is represented by a pair which contains the name of the slot and a sequence of words considered as the value to be assigned to the slot. The slot name represents the meaning of the sequence of words. The proposed annotation scheme considers 60 slots listed below.



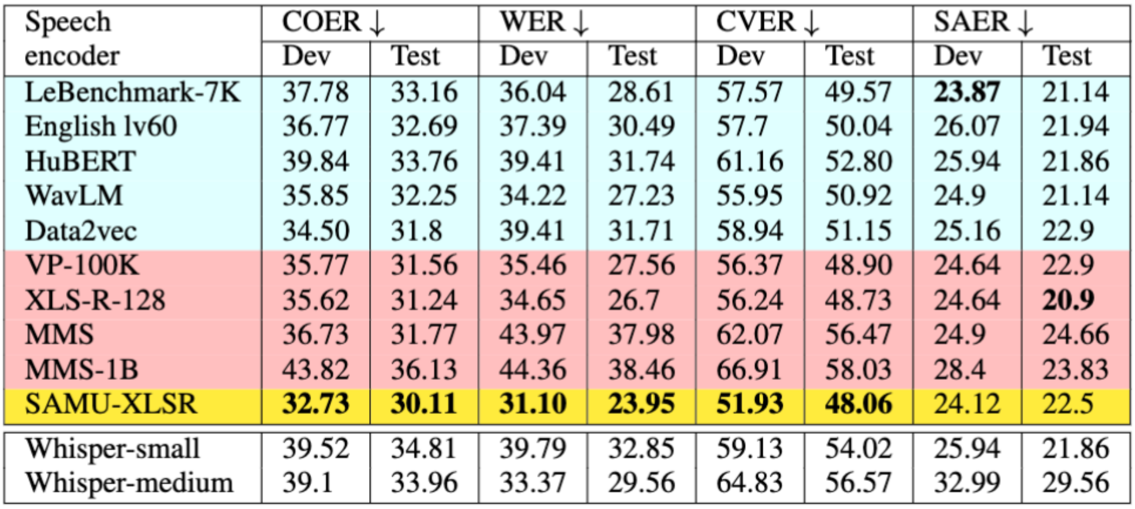
- Age
- Age_request
- Age_ticket
- An
- Answer
- Arrival_time
- Card_type
- Card_price
- City_name_arrival
- City_name_departure
- City_name_before
- City_name_direction
- Class_number
- Class_type
- Command_task
- Comparative_age
- Comparative_distance
- Comparative_price
- Comparative_time
- Coreference_city
- Coreference_departure
- Date
- Day
- Departure_time
- Discount_value
- Discount_percent
- Duration_request
- Duration
- Existence
- Existence_request
- Hour_request
- Money_exchange
- Month
- Negation
- Number
- Number_of_train
- Number_request
- Object
- Option
- Other_transport
- Part_price
- Part_time
- Period_day
- Period_year
- Person_name
- Price_request
- Rank
- Reference_object
- Reference_person
- Reference_time
- Relative_day
- Relative_time
- State
- Tarif
- Task
- Ticket_number
- Ticket_price
- Ticket_type
- Time
- Train_type

***Figure 3** Semantic representation for TARIC-SLU dataset*

Performance of SSL Speech Encoders in the TARIC-SLU dataset

We investigated the application of powerful SSL speech encoders for the execution of SLU in the face of challenging circumstances characterized by a scarcity of SLU training data and the low-resource characteristics of the targeted Tunisian dialect. We used different types of speech encoders.

We considered monolingual SSL speech encoders (French: Wav2vec 2.0 LeBenchmark-7K; English: wav2vec 2.0 LV60, HuBert, wavLM and data2vec) as well as cross-lingual SSL speech encoders (wav2vec 2.0 VP-100K, XLS-R-128, MMS, MMS-1B, SAMU-XLSR).



Speech encoder	COER ↓		WER ↓		CVER ↓		SAER ↓	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
LeBenchmark-7K	37.78	33.16	36.04	28.61	57.57	49.57	**23.87**	21.14
English lv60	36.77	32.69	37.39	30.49	57.7	50.04	26.07	21.94
HuBERT	39.84	33.76	39.41	31.74	61.16	52.80	25.94	21.86
WavLM	35.85	32.25	34.22	27.23	55.95	50.92	24.9	21.14
Data2vec	34.50	31.8	39.41	31.71	58.94	51.15	25.16	22.9
VP-100K	35.77	31.56	35.46	27.56	56.37	48.90	24.64	22.9
XLS-R-128	35.62	31.24	34.65	26.7	56.24	48.73	24.64	**20.9**
MMS	36.73	31.77	43.97	37.98	62.07	56.47	24.9	24.66
MMS-1B	43.82	36.13	44.36	38.46	66.91	58.03	28.4	23.83
SAMU-XLSR	**32.73**	**30.11**	**31.10**	**23.95**	**51.93**	**48.06**	24.12	22.5
Whisper-small	39.52	34.81	39.79	32.85	59.13	54.02	25.94	21.86
Whisper-medium	39.1	33.96	33.37	29.56	64.83	56.57	32.99	29.56

Table 11 *Detailed comparison of speech encoders performance across various SLU tasks. Results reported in WER for ASR transcription, in COER and CVER for slot filling detection and in SAER for speech act classification*

The results are shown in Table 10. The first part of the table shows results by using SSL models, and the second part shows results by using Whisper models. The first part is represented by different colors : the blue color for monolingual models, the pink color for cross-lingual models, and the yellow part for the cross-lingual SSL model semantically refined by an out-of-domain multimodal teacher-student training.

Our results presented in Table 10 underscore the efficacy of SSL pre-trained speech encoders in low-resource conditions, with notable success observed when employing the SAMU-XLSR model. Specifically, the combination of multilingual SSL and teacher-student multimodality training within this model emerged as the optimal approach, yielding superior overall results.

8. Speech-to-text translation

Speech-to-text translation has undergone rapid advancements during the SELMA project. To gain insights into the current state-of-the-art, the research community annually conducts an international evaluation campaign within the framework of the International Conference on Spoken Language Translation (IWSLT). The SELMA project, with a keen interest in knowledge transfer from high-resourced languages to low-resourced ones, has consistently participated by organizing a shared track in the IWSLT campaign for three consecutive years (2022, 2023, and 2024).

8.1. Organization of a shared task on low-resourced language speech translation in IWSLT 2022, 2023 and 2024

Tamasheq, a variety of Tuareg within the Berber macro-language, is spoken by nomadic tribes across North Africa in countries such as Algeria, Mali, Niger, and Burkina Faso, with approximately 500,000 native speakers predominantly in Mali and Niger. This task focuses on translating spoken Tamasheq into written French. The organizers (LIA partner) freely provide nearly 20 hours of spoken Tamasheq with French translations. A significant challenge arises from the absence of Tamasheq transcriptions due to its oral tradition.

The provided corpus consists of radio recordings from Studio Kalangou (https://www.studiokalangou.org) translated into French, totaling 17 hours of clean Tamasheq speech. An extended version of 19 hours includes 2 additional hours flagged by annotators as potentially noisy. Both datasets share identical validation and test sets. For detailed information, refer to (Boito, M. Z., Bougares, F., Barbier, F., Gahbiche, S., Barrault, L., Rouvier, M., & Estève, Y. (2022, June). Speech Resources in the Tamasheq Language. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 2066-2071).

Supplementing the 17 hours of Tamasheq audio aligned with French translations, and considering recent advancements in self-supervised speech processing models, participants are also given unlabeled raw audio data in Tamasheq and four other languages spoken in Niger: French (116 hours), Fulfulde (114 hours), Hausa (105 hours), Tamasheq (234 hours), and Zarma (100 hours), sourced from Studio Kalangou and Studio Tamani (https://www.studiotamani.org).

Each year, an additional separate test set is provided.

8.2. Neural architecture

Our best results for this task were achieved by leveraging an encoder-decoder approach. We employ a SAMU-XLSR model as the speech encoder, as detailed in section 7.2. For the decoder, we utilize the mBART (Liu et al, 2020) model released by Meta.

8.3. Performance

The table below presents the results obtained for this task from 2022 to 2023. The IWSLT 2024 edition is currently underway.

No.	Model	dev	test
1	IWSLT2022-tamasheq-only + Transformer decoder	7.63	5.83
2	IWSLT2022-tamasheq-only + mBART decoder	9.46	7.4
3	SAMU-IWSLT2022-tamasheq-only + mBART decoder	12.6	9.7
4	SAMU-XLSR(53) + mBART decoder	12.5	7.9
5	SAMU-XLSR(60) + mBART decoder	19.1	14.2
6	SAMU-XLSR(100) + mBART decoder	19.3	13.5
7	SAMU-XLSR(100) + mBART decoder (IWLST23 best setup)	21.4	16.5

***Table 12** BLEU results for speech-to-text translation for the task Tamasheq to French*

This table illustrates the rapid evolution of the state-of-the-art over the past two years. It demonstrates the ability to extract the general semantics of utterances from a spoken language with minimal translated data. Further advancements are anticipated in the coming months.

9. Speech-to-speech translation

Speech-to-speech translation has become increasingly important in today's globalized world, facilitating communication across different languages and cultures. However, current speech-to-speech translation systems often fail to preserve the emotional nuances of the speaker's original message, which can lead to misinterpretation and misunderstandings. Emotion is a critical aspect of human communication and preserving it in speech-to-speech translation can greatly enhance the authenticity and effectiveness of cross-cultural interactions, fostering deeper understanding and connection between individuals regardless of languages barriers.

We have proposed a method for speech-to-speech emotion-preserving translation that operates at the level of discrete speech units. Our approach relies on the use of multilingual emotion embedding that can capture affective information in a language-independent manner.

9.1. Learning Multilingual Expressive Speech Representation

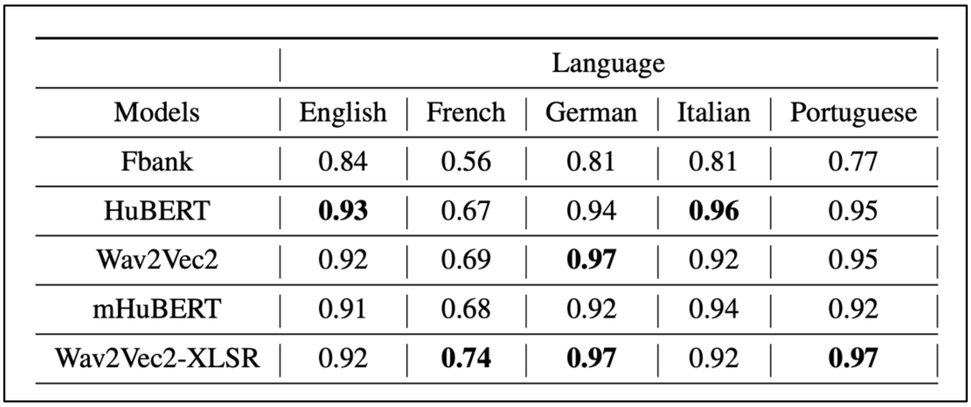
In order to create representations of emotion, a distinct encoder was trained for emotion recognition tasks within a multilingual framework. Our proposed architecture consists of an upstream encoder, Wav2Vec2-XLSR (conneau et al, 2020), that has been pre-trained through semi-supervised learning, a bottleneck layer, and a dense layer with a softmax activation that returns probabilities for each of the emotion classes. Following the approach outlined in (wang et al, 2022), both the CNN and Transformer modules of the Wav2Vec2-XLSR model were fine-tuned during the downstream model training process. In a subsequent step, we employed the previously trained model as an encoder to generate a continuous vectorial representation of dimension 96 for each utterance. This process was accomplished via the bottleneck layer, followed by temporal pooling to reduce the information along the entire speech sequence into a single vector representation.

We experimented with multiple SSL models as encoders on different datasets annotated with emotion labels:

- IEMOCAP, an acted, multimodal and multispeaker database that contains 12 hours of speech from 10 English speakers (5F, 5M);

- CREMA-D a crowd-sourced emotional multimodal acted database containing 7,442 English clips from 91 speakers (43F, 48M). These clips were recorded by actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified);
- ESD, a multilingual emotional database. It consists of 350 parallel utterances spoken by 10 native English and 10 native Chinese speakers (10F, 10M). We only consider the English part;
- Synpaflex, an expressive French audiobooks database. It contains 87 hours of speech, recorded by a single speaker. The subset annotated with emotional labels contains 8,751 clips;
- Oreau, a French emotional speech database containing 502 utterances from 32 non-professional actors (7F, 25M);
- EmoDB, a German emotional speech database containing a total of 535 utterances from 10 professional speakers (5F, 5F);
- EMOVO, an Italian speech database containing 14 sentences over seven different emotions from 6 professional actors (3F, 3M);
- emoUERJ,a Portuguese emotional speech database. It contains 377 utterances from 4 professional speakers (3F, 3F).

Our results in terms of emotion classification accuracy are reported in the following table:

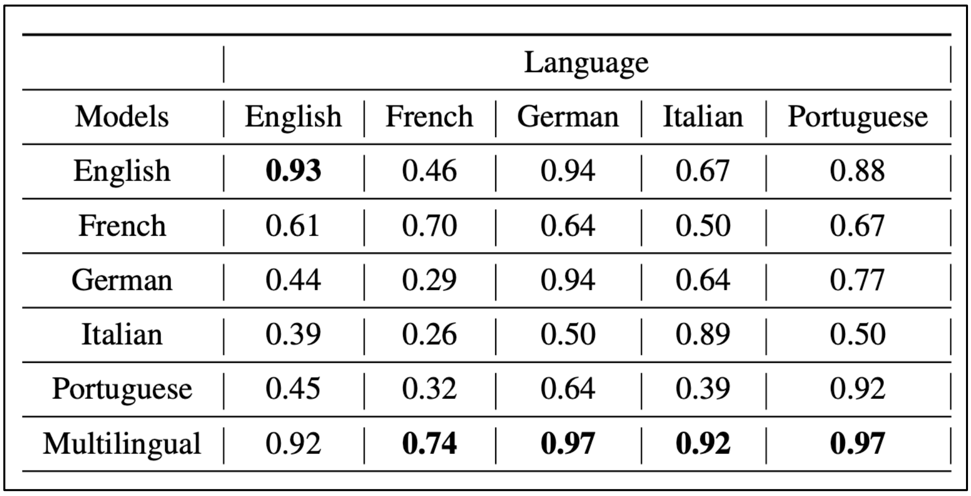


	Language				
Models	English	French	German	Italian	Portuguese
Fbank	0.84	0.56	0.81	0.81	0.77
HuBERT	**0.93**	0.67	0.94	**0.96**	0.95
Wav2Vec2	0.92	0.69	**0.97**	0.92	0.95
mHuBERT	0.91	0.68	0.92	0.94	0.92
Wav2Vec2-XLSR	0.92	**0.74**	**0.97**	0.92	**0.97**

Table 13 *Results from systems trained on multilingual emotion recognition task. Average accuracy (%) ibtained for the different models and languages*

We can see that Wav2Vec2-XLSR outperforms or performs comparably well to other architectures for French, German, and Portuguese, while the difference in performance is marginal for English. Based on this observation and our focus on English and French languages, we choose Wav2Vec2-XLSR as the default architecture for subsequent experiments.

Then, we compare the performance got by these models when trained in monolingual data or multilingual data. The results, in terms of emotion classification accuracy are reported in the following table:



	Language				
Models	English	French	German	Italian	Portuguese
English	**0.93**	0.46	0.94	0.67	0.88
French	0.61	0.70	0.64	0.50	0.67
German	0.44	0.29	0.94	0.64	0.77
Italian	0.39	0.26	0.50	0.89	0.50
Portuguese	0.45	0.32	0.64	0.39	0.92
Multilingual	0.92	**0.74**	**0.97**	**0.92**	**0.97**

Table 14 *Results from systems trained on cross-lingual emotion recognition task. Average accuracy (%) obtained by training the system independently for each language and evaluate its performance on all available languages. For each model we use the Wav2Vec2-XLSR as the default architecture*

The results indicate that the multilingual setup outperforms the monolingual one for all selected European languages, except English by a very minor margin (0.92% of accuracy instead of 0.93%).

This highlights the potential benefits of incorporating multilingual data during training, particularly when the amount of available data for a specific language is very limited.

Next, we conducted an analysis of the embeddings generated by the emotion encoder on a subset of the test corpus. We randomly selected four speakers for each available language and sampled one utterance for each of the emotional labels considered.

Specifically, we applied a k-means clustering with k=4 (the number of emotion labels) on top of the emotion embeddings.

Finally, we present a two-dimensional visualization of the emotion embedding space utilizing the t-SNE algorithm in the following figure:

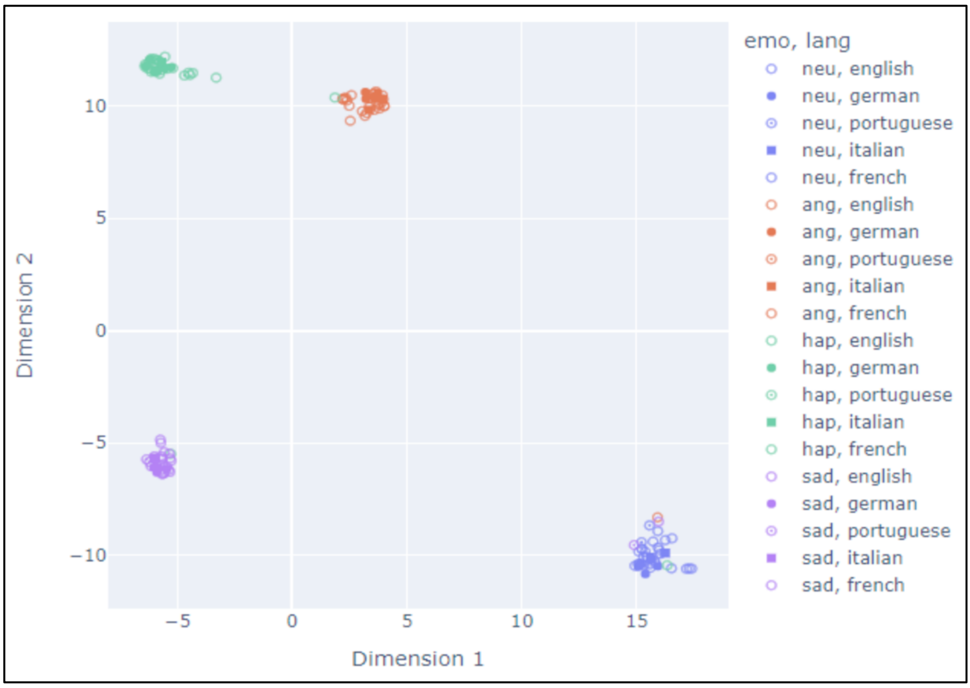


Figure 4 *Visualization of emotion embedding using t-SNE algorithm. In the visualization, each data point corresponds to a distinct utterance, where the emotion label is depicted by distinct colors and the language is represented via various shapes*

This visualization provides a useful representation of the high-dimensional multilingual emotion embedding space and can provide a better understanding of the relationships between different emotions and languages in this space. The observed embeddings exhibit clear separation into four distinct clusters, each corresponding to an individual emotional class, irrespective of the language.

These embeddings were later employed to condition both the Pitch Predictor model and a Duration Predictor model presented in the next sections.

9.2. Speech-to-speech translation preserving expressivity

Recently, a textless direct speech-to-speech translation (S2ST) approach has been proposed that is based on the use of discrete speech units (Lee et al, 2022). Such an approach is particularly interesting to translate from an unwritten language and/or to an unwritten language. It is also noticeable that it is very efficient to process speech-to-speech translation for languages for which a written form exists(Lee et al, 2022, Lee et al,2022). Inspired by this approach, we implemented a solution for speech-to-speech translation preserving expressivity.)

Our speech-to-speech translation framework does not require parallel speech data for speaker and expressivity modeling, enables the translation of speech while maintaining the inherent expressive content, and can generate speech in the target language with multiple voices.

The proposed framework can be decomposed into two parts.

First, a speech-to-unit translation model, composed of a speech encoder and an acoustic decoder. In order to capture the linguistic content, particularly pseudo-phonetic information present in speech, we employ a pre-trained self-supervised learning (SSL) model to extract raw speech features from the audio signal, namely multilingual HuBERT (mHuBERT) for English (target) and Wav2Vec 2.0 for French (source).

Wav2Vec 2.0 and mHuBERT models are pre-trained in a self-supervised manner and produce continuous representations for every 20-ms frame. To extract the sequence of speech units, a k-means clustering is applied to the raw speech features and the learned K cluster centroids are used to transform audio into a sequence of cluster indices at every 20ms of the input audio signal.

Secondly, a unit-to-speech synthesizer, composed of an emotion encoder, a speaker encoder, a duration predictor, a pitch predictor and a speech vocoder.

The overall architecture is illustrated in the following figure:

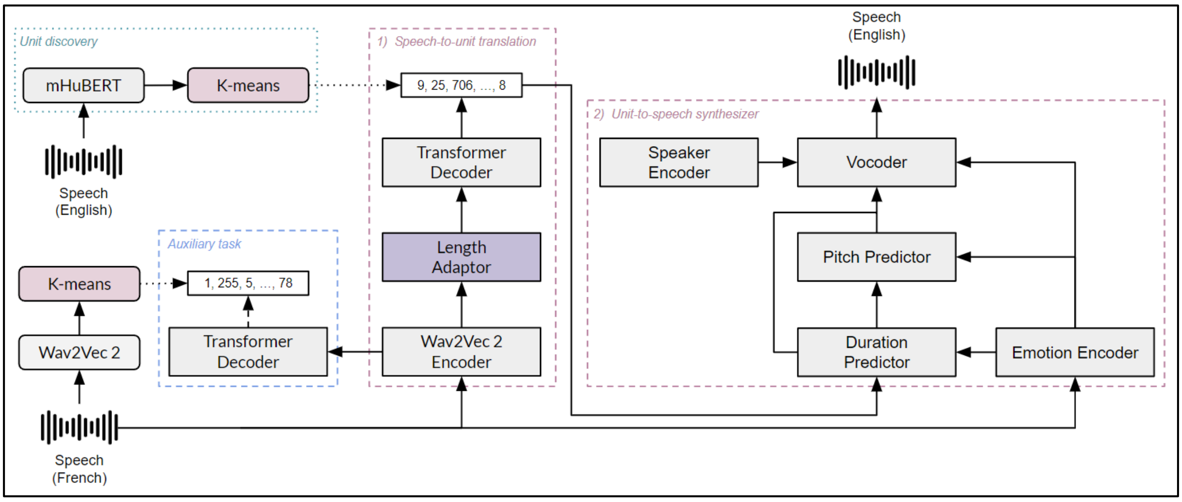


Figure 5 *Illustration of our proposed speech-to-speech translation model. First, the input speech is translated into a sequence of discrete units by the speech-to-unit translation model (1). Next, we predict duration and F0 before feeding them to a unit-to-speech model (2). Duration Predictor, Pitch Predictor, and the unit-to-speech model are conditioned by the emotion embedding extracted from the source speech by the emotion encoder. The speaker is encoded using a 1-hot vector directly in the unit-to-speech model.*

By utilizing this method, we aim to improve the accuracy of speech-to-speech translation models in capturing the linguistic content of the target speech without being influenced by the speaker's prosodic features.

Following (Polyak et al, 2021), we use the HiFi-GAN neural vocoder (Kong el al, 2020) to synthesize speech. HiFiGAN is a generative adversarial network (GAN) that consists of one generator and a set of discriminators.The generator is a fully convolutional neural network. Inspired by (Kreuk el al, 2022), we adapted the generator architecture to take as input a sequence of discrete-unit inflated using the predicted durations, predicted F0, emotion-embedding, and a speaker-embedding. Before feeding the above features into the model, we concatenate them along the temporal axis. The sample rates of unit sequence and F0 are matched by means of linear interpolation, while the speaker-embedding and emotion-label are replicated along the temporal axis.

Regarding the set of discriminators, the model is composed of two modules: a Multi-Scale Discriminators and a Multi-Period Discriminators. The first type operates on different sizes of sliding windows over the input signal, while the latter samples the signal at different periods.

Experiments

We use the SpeechMatrix corpora for training and evaluating our **speech-to-unit translation (S2UT) model.** SpeechMatrix consists of 126 language pairs with a total of 418 thousand hours of speech from European Parliament recordings. In this study, only French-to-English language pairs were considered, yielding a 1,507-hour train set.

In addition to the mined speech-to-speech data for training purposes, we extend our evaluation by leveraging labeled public speech datasets obtained from two distinct corpora that cover various domains. First, Europarl-ST (EPST), a multilingual corpus containing paired audio-text samples built from recordings of debates from the European Parliament, containing 72 translation directions in 9 languages, including French to English direction.

The second dataset is FLEURS(Conneau et al, 2023). Derived from from FLoRes, FLEURS is an extension that introduces speech recordings for these translated texts, resulting in a collection of speech-to-speech data comprising French to English direction. FLEURS texts are from English Wikipedia.

The unit-to-speech (U2S) model is separately trained from S2UT model. To train the U2S system for English language, we combine the LJSpeech dataset and the ESD dataset. The LJSpeech dataset contains 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books, with a total duration of approximately 24 hours. ESD is a multilingual emotional database, consisting of 350 parallel utterances spoken by 10 native English and 10 native Chinese speakers (10F, 10M). In this study, we only consider the English part.

During training, we extract a validation set from SpeechMatrix of about 1000 samples which are not in the test set. FLEURS validation set is derived from its validation samples.

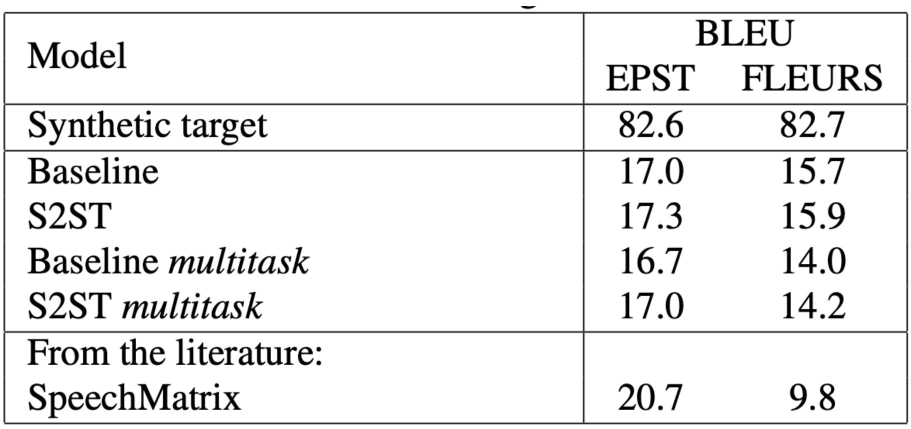
To compute evaluation scores, we consider only the source speech and target texts,

To assess the effectiveness of our proposed approach, we built a **Baseline** model which is composed of a speech-to-unit translation (S2UT) module and a unit-to-speech (U2S) module. We conduct an analysis by systematically excluding the emotion encoder and pitch predictor from the U2S module. This enables us to quantify the impact and measure the benefits of their inclusion in the overall system.

In addition to the Baseline and our S2ST model, we also incorporated an English text-to-speech (TTS) model (Kim el al, 2021) into our subjective evaluation. The TTS model was trained on the identical dataset utilized for training the U2S module. Its inclusion serves to assess the overall quality of the synthesized speech generated by the TTS model in comparison to our proposed system and to evaluate the effectiveness of expressivity transfer achieved by our proposed system in contrast to the TTS model.

Recent work in speech-to-speech translation suggests evaluating translation quality using the BLEU score. We start by using an ASR model to compute the transcriptions of the generated speech. In order to obtain comparable results, we use the same open-source ASR model as in (Duquenne et al, 2022). Then, **we compute BLEU score of the ASR decoded text with respect to the reference translations**.

Results are presented in the following table:



Model	BLEU	
	EPST	FLEURS
Synthetic target	82.6	82.7
Baseline	17.0	15.7
S2ST	17.3	15.9
Baseline *multitask*	16.7	14.0
S2ST *multitask*	17.0	14.2
From the literature:		
SpeechMatrix	20.7	9.8

Table 15 *BLEU scores on EPST and FLEURS test sets by S2ST models with different settings*

First, we compare the proposed S2ST model to the Baseline. We can see that our S2ST model outperforms the Baseline by 0.3 BLEU on EPST and by 0.2 BLEU on FLEURS, indicating that our approach performs similar or slightly better in terms of translation performance.

We also note that SpeechMatrix achieves an improvement of 3.4 BLEU over the proposed S2ST model on EPST, however, on FLEURS our approach outperforms SpeechMatrix by 6.1 BLEU leading to an average improvement of 1.3 BLEU. The gap of performance on the FLEURS test set can be attributed in part to the fact that we use an encoder pre-trained on 7000 hours of speech covering multiple domains compared to SpeechMatrix encoder trained only on European Parliament recording.

Secondly, we explore multitask learning by incorporating an auxiliary task to the Baseline and S2ST model. In our experimental setup, we observe a decline in performance for both the Baseline and the S2ST model when employing multitask learning. Specifically, the S2ST model yields a performance of (17 vs. 17.3) on EPST and (14.2 vs. 15.9) on FLEURS. This suggests that our encoder does not provide significant benefits to the auxiliary task. Nonetheless, our approach still outperforms the Baseline system for both setups, indicating the effectiveness of our proposed approach.

In addition to measuring the translation quality via an objective metric, we conduct **human listening tests to assess perceptual responses** of expressivity transfer from recordings generated by our S2ST model.

We asked 33 people to evaluate two sets of tasks online. A detailed description of the tasks was provided to all evaluators, who had unlimited time to evaluate audio stimuli. Each task was organized similarly, consisting of one pre-trial (excluded from this analysis) and four trials. Each trial contained three synthesized speech recordings produced by Baseline, TTS, and our S2ST framework.

After listening to each recording, evaluators provided an opinion score on a scale of 1 to 5, where 1 is `Poor' and 5 is `Excellent'. The first task was a **Mean Opinion Score (MOS)** where evaluators judged the quality of the synthesized speech. The second task was a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA), where evaluators listened to a reference (natural, spoken French) and then judged the expressiveness of the English-translated synthesized speech.

Results are presented in the following figure:

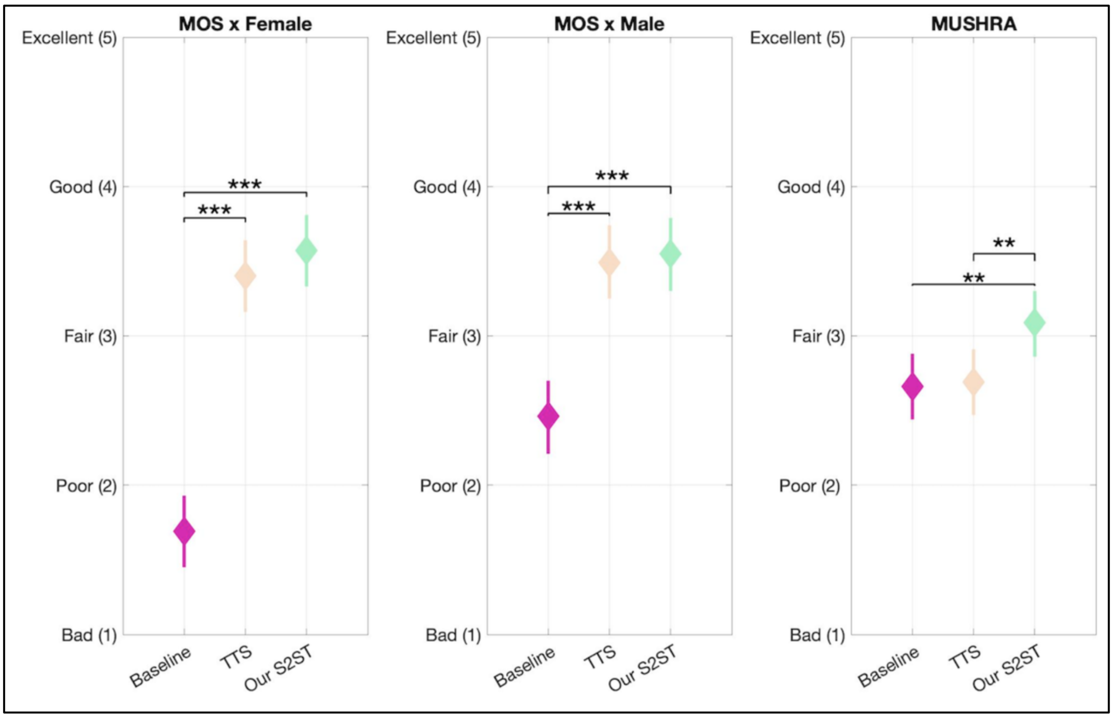


Table 16** MOS (Left-Middle) and MUSHRA (Right) task results. Diamonds and vertical lines represent mean and critical intervals. {**,} represent p<{0.01,0.001}*

Separate linear mixed effects models were used to evaluate MOS and MUSHRA task responses. Using the R-package *lme4*, opinion responses were entered as response variables. Synthesized speech system (3-levels) and speaker sex (2-levels) were entered as fixed factors and participant was entered as a random factor. Chi-squared ($\chi^2_{d,N}$) tests were used to report p-values with *d* degrees of freedom and *N* samples, i.e., there were *N=486* responses, *d*=2 speech systems, and *d* = 1 speaker sexes. Main effects were reported for task, response, and their interactions with speaker. Estimated marginal means were used to conduct pairwise comparisons, where *X* ± *Y* represents mean and standard error, respectively.

The results of the MOS task revealed significant main effects on system $\chi^2_{2,486}$ = 284.17 and speaker sex $\chi^2_{1,486}$ = 11.25, as well as their interaction $\chi^2_{2,486}$ = 18.66, p < 0.001. Pairwise comparisons showed that the quality of recordings generated by the Baseline system (2.07 ± 0.1) had significantly lower opinion scores in comparison to those generated by TTS (3.45 ±

0.1) and our S2ST model systems (3.56 ± 0.1), $p < 0.001$. In comparison to female speech recordings (2.89 ± 0.09), male speech recordings (3.17 ± 0.09) had significantly increased scores, $p < 0.001$, however, these effects were localized to the Baseline system.

There are several takeaways from our subjective evaluations. First our S2ST framework produced speech recording that were perceived to have higher quality in comparison to those produced by the Baseline system. Next it outperformed both Baseline and TTS systems in terms of producing recordings that conveyed speaker expressivity.

The experimental results have demonstrated the superior expressivity transfer achieved by our method compared to state-of-the-art systems, highlighting its effectiveness. Moreover, our speech-to-speech translation framework has produced speech recordings that were perceived by humans to have higher quality in terms of conveying speaker expressivity, surpassing both our speech-to-speech Baseline and text-to-speech systems. Importantly, we have maintained the translation quality at a level similar to that of state-of-the-art textless speech-to-speech translation systems.

10 Taking into Account the user Feedback: the M-PHANTOM model

Neural-based ASR systems struggle to correctly spell named entities or domain-specific words because of their scarcity during training of the underlying neural network. The situation becomes direr when one considers zero-shot scenarios, i.e., the speech contains words that were never seen during training. Current solutions consider using external knowledge from pre-defined databases of named entities, but few explore the setting where the user may give feedback by correcting the misspellings on the automatically generated transcripts. Moreover, approaches that apply further training to the model weights given external knowledge are hard to deploy on industrial scales, and thus preference should be given to solutions that rely on contextual biasing the transcription process, like prompting for LLMs.

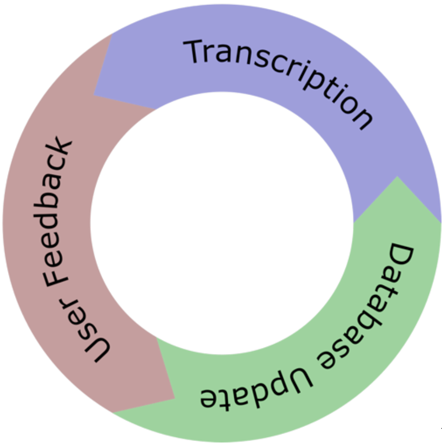
The study we made allowed the development of an extension to CB-Whisper (Li et al., 2024) that considers user feedback. CB-Whisper was recently proposed as an ASR system that prompted a Whisper (Radford et al., 2022) model with a biasing list of named entities, so-called keywords, thus conditioning its transcription process to adequately spell those words. The keywords to be used are retrieved from an external database of TTS-generated speech for the keywords using an open-vocabulary keyword-spotting (OV-KWS) classifier. To the best of our knowledge, our extension, named M-PHANTOM, is the first to take the human in the loop and uses the human-corrected transcripts to identify new keywords to be added to the database together with corresponding audio segments extracted directly from natural speech, thus circumventing pitfalls common to TTS-generated speech.

10.1. Background

Contextual biasing refers to injecting prior knowledge into ASR systems to make them achieve higher recalls for named entities or domain-specific words without need of retraining. There has been a large research effort in developing model-based solutions, which incorporate biasing components in the models, usually based on attention mechanisms, to condition the model outputs on the biasing context (Xu et al., 2023, Naowarat et al., 2023). Several other works take distinct approaches, for instance, SALM (Chen et al., 2023) is a fusion of an LLM with a speech module so that the ASR system can consider lists of named entities through in-context learning. Another work (Wang et al., 2023) proposed a TPU-friendly implementation of pattern-matching based biasing that may be used during decoding of the ASR model.

Whisper (Radford et al., 2022) is a collection of multilingual ASR encoder-decoder models proposed by OpenAI whose generation can be biased with keywords provided to its prompt. CB-Whisper builds off Whisper and resorts to an additional OV-KWS classifier to retrieve keywords from an external database to form a restricted biasing list. However, the external database relies on TTS-generated audios for every keyword, which is bound to be poor performant for graphemes that have non-trivial phonetic transcriptions, like abbreviations, acronyms, acronym-abbreviations, to name a few (Burkhardt et al., 2016).

10.2. M-PHANTOM Overview



wh

Figure 6 *Depiction of the feedback loop of M-PHANTOM*

M-PHANTOM appears as a natural extension to CB-Whisper that takes place on settings where the ASR system may receive user feedback through corrections of the transcripts. The availability of user feedback not only allows to further extend the database with new keywords identified from the misspelled words, but also allows to obtain natural-speech audios for the keywords by segmenting the source audio that the user inputs, and thus circumvents the disadvantages of using TTS-generated speech for the keywords.

M-PHANTOM contains a Whisper model for speech recognition and an OV-KWS model. The latter is a binary classifier based on convolution neural networks that takes the dense features of the target utterance and of each keyword pairwise and determines whether a given keyword is in the audio or not. The dense features are obtained using the Whisper-medium encoder from a given audio. The dense features of each keyword are stored in a database, together with the keywords themselves and other relevant metadata, for instance, the audios from which those features were derived.

The human-computer interaction workflow may be described in three main stages: transcription, user feedback and database update, which are depicted in figure above.

- *Transcription*: the user provides an audio that is subsequently transcribed using CB-Whisper and the current state of the keyword database.
- *User feedback*: the user corrects the misspellings on the automatically generated transcript.
- *Database update*: the system identifies and promotes the misspelled words to keywords, obtains speech for each keyword by directly segmenting the source audio, and adds the new keywords to the database together with their dense features.

The database update requires access to word-level timestamps to segment the input audio and obtain speech for each keyword. To that end, we make use of Stable-Whisper, which forced-aligns the input audio and the human-corrected transcript. To identify the misspelled words, the original and human-corrected transcripts are aligned with the Needleman-Wunsh algorithm at character-level. The words from the human-corrected transcript with mismatched characters are promoted to keywords. Subsequent transcriptions will use the newly updated keyword database.

10.3. Revisiting OV-KWS

M-PHANTOM presents the possibility of applying OV-KWS to features derived from natural-speech keyword audios. Although it bypasses several limitations of using TTS models for generating speech for the keywords, it comes with its own challenges for building OV-KWS models that can generalize to different speakers, difficult recording conditions, etc... For this reason, we conducted several experiments to build robust OV-KWS models which are more adequate to deal with natural speech and to assess the generalization capabilities to unseen modalities (TTS-generated or natural-speech keyword audios) and unseen languages.

Datasets:

We used two datasets to set up the training data for the experiments:

- Aishell-1 (Bu et al., 2017), which is a Chinese ASR dataset whose training split contains 120098 utterances with a total of around 150 hours of speech.
- Multilingual Librispeech (MLS) (Pratap et al., 2020), which is a multilingual ASR dataset. Since the number of hours per language is imbalanced, we selected 25 hours from six languages (English, French, German, Polish, Portuguese and Spanish) with 35356 utterances. The selection of the utterances aimed at including as many different speakers as possible.

These two ASR datasets needed to be adapted for keyword-spotting, and the resulting datasets were called Aishell-KWS and MLS-KWS. The first step was to construct a vocabulary of keywords. For Aishell-KWS, 20000 keywords were selected from an existing lexicon provided together with the data. The same number of keywords was used for each one of the languages of MLS-KWS, which were sampled from all whitespace-separated words in the transcripts. Aishell-KWS and MLS-KWS contain 2397055 and 1011956 available positive examples, respectively. A positive example consists of a paired utterance and keyword in the same language such that the latter is present in the former, otherwise called negative example. For validation and final evaluation purposes, two datasets were used:

- Aishell hotword dev and test subsets (Shi et al., 2023) in Chinese, hereby denoted Aishell-dev and Aishell-test sets
- ACL6060 dev and test subsets (Salesky et al., 2023) in English, hereby denoted ACL6060-dev and ACL6060-test sets

The speech for the keywords may be either TTS-generated or extracted from natural speech. Synthetic keyword audios are generated using edge-tts with randomly sampled synthetic voices. Natural-speech keyword audios are segments from the utterances where those words are spoken. Stable-Whisper was used to obtain the required word-level timestamps for every occurrence of a particular keyword in the dataset, but only the alignment with highest confidence was kept. Every training, validation or test dataset may be in either one of both modalities, either TTS-generated or natural speech for the keywords, which shall be disambiguated with the use of TTS or NAT, respectively.

Evaluation Metrics:

The metric used to evaluate the performance of the classifiers for OV-KWS is the F1 score, which is the harmonic mean between the precision and recall.

Experiments Setup:

The dense features for both the utterances and keywords are extracted from the last 12 encoder layers of Whisper-medium. The input to the classifier for OV-KWS are cosine similarity matrices between the dense features of the keywords and the ones of the utterance and were resized to (150, 750). The OV-KWS model is a binary CNN classifier identical to a ResNet-50 (He et al., 2016) for image classification. Every epoch contains 4 examples from each utterance, one positive and three negative. Among the negative examples, one is chosen randomly and the other two are chosen to be in the same language of the utterance and lexicographically close to the positive example.

We conducted several experiments to assess the impact of using keyword features derived solely from TTS-generated audios or natural-speech segments or from a mixture of both and the generalization capabilities of the classifier to a different modality or unseen languages. When using a mixture (TTS + NAT), we randomly chose whether to use keyword features derived from TTS-generated or natural speech audio with equal probability. During training, checkpoints were validated against the dev sets that share the same language and modality used in training. For instance, training with Aishell-KWS and validation with Aishell-dev TTS. For each experiment, the checkpoint that achieved a higher averaged F1 score among these used dev sets was kept for further evaluation.

Experimental Results:

The results from the experiments are presented in the following table:

Experiments		**Aishell-test (zh)**		**ACL6060-test (en)**	
		TTS	**NAT**	**TTS**	**NAT**
Aishell-KWS (zh)	**TTS**	**89 ± 3** (-00%)	10 ± 1 (-89%)	53 ± 5 (-20%)	10 ± 3 (-83%)
	NAT	6 ± 5 (-93%)	**89 ± 3** (-00%)	5 ± 4 (-92%)	44 ± 6 (-24%)
	TTS + NAT	88 ± 3 (-01%)	**89 ± 4** (-00%)	52 ± 5 (-21%)	44 ± 6 (-24%)
MLS-KWS (en + others)	***TTS***	41 ± 6 (-54%)	2 ± 1 (-98%)	**65 ± 4** (-02%)	4 ± 1 (-93%)
	NAT	2 ± 2 (-97%)	75 ± 4 (-16%)	2 ± 2 (-97%)	**56 ± 4** (-03%)
	TTS + NAT	71 ± 5 (-20%)	68 ± 5 (-24%)	**65 ± 4** (-02%)	53 ± 4 (-09%)

Table 17 *OV-KWS (F1%) results on Aishell-test and ACL6060-test with both modalities TTS and NAT. Each score contains a 95% confidence interval computed using bootstraping. The best scores are identified in bold and the grey numbers indicate the percentage drop relative to the best score in the same test set*

The best results are achieved by models that are evaluated in the same domain (the same language and modality are used in training). When considering the same language in evaluation as the one used in training, TTS and NAT experiments exhibit lower-than-anticipated results when tested on NAT and TTS datasets, respectively. The decline in performance, in the most extreme case (MLS-KWS NAT assessed on ACL6060-test TTS), reaches 97%. This underscores the incapacity of the OV-KWS classifier to generalize to keyword features derived from natural speech solely based on training with those from TTS-generated speech, and vice versa. Combining both TTS and NAT modalities during training mitigates this issue, maintaining comparable in-domain performance for TTS and NAT experiments separately, within statistical significance.

When assessing test languages not encountered during training, the F1 scores of TTS + NAT experiments, though not as robust as in-domain evaluations, experience a maximum performance drop of only 24%. Training with Chinese data does not notably enhance generalization to English across experiments. In contrast, when using a multilingual dataset for training, the NAT demonstrates superior generalization performance to Chinese, if evaluated in the same modality, whereas training with both modalities achieves the best result when evaluated on the TTS modality.

Discussion:

The findings indicate that the OV-KWS classifier struggles to generalize to a modality that it has not encountered during training. However, this limitation can be effectively remedied by incorporating a mixture of both modalities in the training. This discovery implies that employing natural-speech audio for keywords or a combination of modalities are the most suitable approaches for training a classifier for OV-KWS intended for integration into M-PHANTOM. Moreover, there is indication that generalizing to unseen languages could be enhanced with a multilingual dataset, although the performance appears to be influenced significantly by the modality utilized during training.

Additional work not included in this report:

Further experiments with Domain Adversarial Neural Networks (Matsuura et al., 2020) were performed to improve the generalization capabilities of the OV-KWS classifier to different modalities and languages. Although not included in this report, their results together with the ones indicated in this section are included in a paper that was submitted to Interspeech 2024. Moreover, we also directly measured the performance of M-PHANTOM on correctly spelling the keywords on the Aishell-test and ACL6060-test sets, but we do not include these results because they contain little to no novelty with respect to the results reported on the original paper of CB-Whisper.

11 Conclusion

The final report summarizes the research conducted during the SELMA project.

Over the 3 years of SELMA, the technologies focused on by the project have evolved rapidly. Research partners have successfully adapted their work to these changes, including pretraining SSL wav2vec2.0 models, utilizing Whisper models, exploring the use of discrete speech units for textless speech translation, and investigating semantic speech encoders.

The technical skills, competencies, and agility of the SELMA project partners have enabled the research presented in this report. Throughout, they prioritized transfer learning and integrating user feedback, as outlined in the initial proposal.

References

A

Ardila et al. 2020 https://arxiv.org/abs/1912.06670

B

Bachman et al., 2019 https://arxiv.org/abs/1906.00910

Baevski et al., 2019 https://arxiv.org/abs/1911.03912

Baevski et al., 2020, https://arxiv.org/abs/2006.11477

Bahdanau et al., 2014 https://arxiv.org/abs/1409.0473

Bänziger et al., 2012 https://pubmed.ncbi.nlm.nih.gov/22081890/

Bisani and Ney, 2004 https://ieeexplore.ieee.org/document/1326009

Boito et al., 2020 https://arxiv.org/abs/1907.12895

Bonneau Maylard et al., 2006 https://aclanthology.org/L06-1385/

Branca-Rosoff et al., 2012 http://cfpp2000.univ-paris3.fr/CFPP2000.pdf

Bu et al., 2017 https://ieeexplore.ieee.org/document/8384449

Burkhardt et al., 2016 http://www.lrec-conf.org/proceedings/lrec2016/summaries/208.html

C

Chan et al., 2016 https://ieeexplore.ieee.org/document/7472621

Chen et al., 2020 https://arxiv.org/abs/2002.05709

Chen et al., 2023 https://arxiv.org/abs/2310.09424

Chung et al., 2019 https://arxiv.org/abs/1904.03240

Chung and Glass, 2020-A https://arxiv.org/abs/2004.05274

Chung and Glass, 2020-B https://arxiv.org/abs/1910.12607

Conneau et al., 2020 https://arxiv.org/abs/2006.13979

Conneau et al, 2023 https://arxiv.org/pdf/2205.12446.pdf

D

Devlin et al., 2018 https://arxiv.org/abs/1810.04805

De Mori, 1997 https://www.elsevier.com/books/spoken-dialogue-with-computers/de-mori/978-0-12-209055-4

Dinarelli et al., 2017 https://hal.archives-ouvertes.fr/hal-01553830v1;

Dinarelli et al., 2020 https://arxiv.org/abs/2002.05955

Duquenne et al, 2022, https://aclanthology.org/2023.acl-long.899.pdf

E

Eshkol-Taravella et al., 2012 https://halshs.archives-ouvertes.fr/halshs-01163053/document

Estève et al., 2010 https://aclanthology.org/L10-1442/

Evain et al., 2021-A https://arxiv.org/abs/2104.11462

Evain et al., 2021-B https://openreview.net/forum?id=TSvj5dmuSd

F

Feng et al., 2022 https://arxiv.org/pdf/2007.01852.pdf

G

Gournay et al., 2018 https://www.researchgate.net/publication/326022359_A_canadian_french_emotional_speech_dataset

Graves et al., 2006 https://dl.acm.org/doi/10.1145/1143844.1143891

Gravier et al. 2012 https://aclanthology.org/L12-1270/

Gururangan et al., 2020 https://arxiv.org/abs/2004.10964

H

He et al., 2016 https://ieeexplore.ieee.org/document/7780459

Hochreiter and Schmidhuber, 1997 https://dl.acm.org/doi/10.1162/neco.1997.9.8.1735

Hsu et al., 2021 Hsu et al., 2021

K

Kalchbrenner et al., 2018 https://arxiv.org/abs/1802.08435

Kawakami et al., 2020 https://arxiv.org/abs/2001.11128

Koehn et al., 2004 https://aclanthology.org/W04-3250/

Khurana et al. 2022 https://arxiv.org/abs/2205.08180

Kim et al, 2020 https://arxiv.org/abs/2005.11129

Kim el al, 2021 https://arxiv.org/pdf/2106.06103.pdf

Kong et al., 2020 https://arxiv.org/abs/2010.05646

Kreuk el al, 2022 https://arxiv.org/pdf/2111.07402.pdf

L

Laperrière et al, 2022 https://arxiv.org/pdf/2210.05291.pdf

Lefèvre et al., 2012 https://hal.archives-ouvertes.fr/hal-01434925

Le Moine et al., 2020 https://arxiv.org/abs/2004.04410

Li et al., 2024, https://arxiv.org/abs/2309.09552

Liu et al., 2019 https://arxiv.org/abs/1910.12638

Liu et al., 2020 https://arxiv.org/pdf/2001.08210.pdf

Lee et al, 2022 https://aclanthology.org/2022.acl-long.235/

Lee et al,2022 https://aclanthology.org/2022.naacl-main.63/

M

Masmoudi el al., 2014 http://www.lrec-conf.org/proceedings/lrec2014/pdf/454_Paper.pdf

Matsuura et al., 2020 https://ojs.aaai.org/index.php/AAAI/article/view/6846

Mdhaffar et al., 2022 https://www.isca-speech.org/archive/pdfs/interspeech_2022/mdhaffar22_interspeech.pdf

Mdhaffar el al., 2024 paper accepted in LREC 2024 (May) (not currently available online)

N

Naowarat et al., 2023 https://www.isca-archive.org/interspeech_2023/naowarat23_interspeech.html

Nguyen et al., 2020 https://hal.archives-ouvertes.fr/hal-02962186

O

ORTOLANG-MPF https://hdl.handle.net/11403/mpf/v3

ORTOLANG-TCOF https://hdl.handle.net/11403/tcof/v2.1

Ott et al., 2019 https://arxiv.org/abs/1904.01038

P

Peters et al., 2018 https://arxiv.org/abs/1802.05365

Peddinti et al., 2015 https://www.isca-speech.org/archive_v0/interspeech_2015/papers/i15_3214.pdf

Pinnis et al., 2014 https://aclanthology.org/L14-1257/

Polyak et al, 2021 https://arxiv.org/pdf/2104.00355.pdf

Povey et al., 2011 https://www.danielpovey.com/files/2011_asru_kaldi.pdf

Povey et al., 2016 https://www.isca-speech.org/archive_v0/Interspeech_2016/pdfs/0595.PDF

Povey et al., 2018 https://www.isca-speech.org/archive_v0/Interspeech_2018/pdfs/1417.pdf

Pratap et al., 2020 https://arxiv.org/abs/2012.03411

Q

Quarteroni et al., 2009 http://www.marcodinarelli.it/NewSite/styles/publications/Interspeech09-Ontology.pdf

R

Radford et al., 2022 https://arxiv.org/abs/2212.04356

Raffel et al., 2019 https://arxiv.org/abs/1910.10683

Ravanelli et al., 2021 https://arxiv.org/abs/2106.04624

Ren et al., 2019 https://arxiv.org/abs/1905.09263

Riviere et al., 2020 https://arxiv.org/abs/2002.02848

Ruder, 2021 https://ruder.io/nlp-benchmarking/

S

Salesky et al., 2021 https://arxiv.org/abs/2102.01757

Salesky et al., 2023 https://aclanthology.org/2023.iwslt-1.2/

Schlangen, 2021 https://arxiv.org/abs/2007.04792

Schneider et al., 2019 https://arxiv.org/abs/1904.05862

Shen et al., 2018 https://arxiv.org/abs/1712.05884

Shi et al., 2023 https://arxiv.org/abs/2308.03266

SLR57, https://www.openslr.org/57/

Snyder, 2015 https://arxiv.org/abs/1510.08484

Song et al., 2019 https://arxiv.org/abs/1910.10387

Stolcke, 2002 http://www.speech.sri.com/projects/srilm/papers/icslp2002-srilm.pdf

T

Torreira et al., 2010 https://hal.archives-ouvertes.fr/hal-00608402

V

Vainer and Dusek, 2020 https://arxiv.org/abs/2008.03802

Vaswani et al., 2017 https://arxiv.org/abs/1706.03762

W

Wang et al., 2020 https://arxiv.org/abs/2001.10603

Wang et al., 2021 https://arxiv.org/abs/2101.00390

Wang et al, 2022 https://arxiv.org/pdf/2111.02735.pdf

Wang et al., 2023 https://arxiv.org/abs/2310.00178

X

Xu et al., 2023 https://www.isca-archive.org/interspeech_2023/xu23d_interspeech.html

Y

Yang et al., 2021 https://arxiv.org/abs/2105.01051