Towards Speech-to-Pictograms Translation

Cécile Macaire, Chloé Dion, Didier Schwab, Benjamin Lecouteux,
Emmanuelle Esperança-Rodier

Towards Speech-to-Pictograms Translation

Cécile Macaire[1], *Chloé Dion*[1], *Didier Schwab*[1], *Benjamin Lecouteux*[1], *Emmanuelle Esperança-Rodier*[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

firstName.lastName@univ-grenoble-alpes.fr

Abstract

The automatic translation of speech into pictogram terms (Speech-to-Pictos) represents a novel NLP task with the potential to enhance communication for individuals with language impairments. Recent research has not explored the adaptation of state-of-the-art methods to this task, despite its significance. In this work, we investigate two approaches: (1) the cascade approach, which combines a speech recognition system with a machine translation system, and (2) the end-to-end approach, which tailors a speech translation system. We compare state-of-the-art architectures trained on an aligned speech-to-pictogram dataset, specially created and released for this study. We conduct an in-depth automatic and human evaluation to analyze their behavior on pictogram translation. The results highlight the cascade approach's ability to generate relevant translations from everyday read speech, while the end-to-end approach achieves competitive results with challenging acoustic data.

Index Terms: automatic speech recognition, machine translation, speech translation, pictograms, augmentative and alternative communication.

1. Introduction

Alternative and Augmentative Communication (AAC) encompasses tools and strategies to facilitate communication for individuals facing language impairment [1]. These disorders affect various language abilities, from speech production to listening, reading, and writing. Genetic diseases, autistic spectrum disorders, and intellectual deficits can be the cause of it. In AAC, pictograms serve to convey messages in everyday life situations. It is a graphic representation associated with a concept (object, person, action, etc.) [2], offering benefits to visualize syntax, manipulate words, and facilitate language access [3]. Moreover, pictograms are widely utilized resources among medical institutes, caregivers, and families [1]. From a social perspective, a 2021 French Red Cross survey identified stress reduction, increased autonomy, and positive well-being for AAC users. However, the study identifies various environmental barriers that limit its use and dissemination. The survey specifically mentions a lack of awareness among caregivers about the potential of AAC and the difficulty of accessing tools (lack of information, training, financial resources, and time).

We believe that implementing speech-to-pictograms translation systems, a task we will refer to as Speech-to-Pictos (S2P), could help address these challenges. As printed in Table 1, the S2P system predicts a sequence of terms (*pictos_tokens*), each associated with a unique ARASAAC [2] pictogram (*arasaac_pictos*), from an audio segment.

[1]https://arasaac.org/world

[2]https://arasaac.org/

Table 1: *Illustration of the S2P task, with the audio segment as input, and output the sequence of terms (pictos_tokens).*

Automatic translation from french speech to a sequence of pictograms was studied for the first time by Vaschalde et al. [4]. Their methodology adapts the Text2Picto system [5] by integrating four modules: an Automatic Speech Recognition (ASR) system, a simplification system, a word sense disambiguation model, and a module to display the sequence of pictograms. The preliminary study does not report automatic or human evaluations. To our knowledge, there exists no other studies exploring the translation of speech into pictograms for French.

Contributions:

1. We introduce two approaches to automatically translate speech into pictogram terms, cascaded and end-to-end.
2. We construct and release two freely available corpora of aligned speech-text-pictograms for this task.
3. We implement and publish state-of-the-art Automatic Speech Recognition (ASR), Machine Translation (MT), and Speech Translation (ST) models adapted to the datasets [3].
4. We present both an automatic and an original human evaluation for the S2P task, which yield competitive results with a cascade approach for our target group.

2. Proposed methods

In this work, we explore both cascade and end-to-end approaches for the Speech-to-Pictos (S2P) task, focusing on state-of-the-art models presented in the following sections.

2.1. Cascade approach

The cascade approach consists of an ASR system and an MT system. The transcription provided by the ASR system is the entry point for the MT system, whose goal is to translate a source language sentence $X = (x_1, ..., x_n)$ (the French transcription) into the target language sentence $Y = (y_1, ..., y_n)$ (the *picto_tokens* sequence). For the S2P task, the target language is the "pictographic language" corresponding to the se-

[3]The code is released at https://github.com/macairececile/speech-to-pictograms

quence of terms (single word, multi-word expression, or entire sentence), each associated with an ARASAAC pictogram.

Automatic speech recognition. Impressive results have been demonstrated in downstream speech tasks with the application of self-supervised learning (SSL) [6, 7]. We consider Wav2Vec 2.0 [8], which learns robust speech representations from a large collection of unlabeled data. The architecture is based on convolutional layers and Transformers [9]. Wav2vec 2.0 is then fine-tuned on labeled data with a Connectionist Temporal Classification (CTC) loss [10]. Recent works have introduced massively multimodal and multilingual models, achieving competitive ASR results without the need for a fine-tuning step. Whisper [11] employs the Transformer architecture [9] trained in a weakly supervised pre-training fashion with 680,000 hours of multilingual labeled data (from 100 languages). SeamlessM4T [4] is a multimodal and multilingual machine translation model covering a hundred languages. The architecture combines a Transformer textual encoder-decoder NLLB, a Conformer-based W2v-BERT 2.0 speech encoder [12], a Text-to-Unit encode-decoder and a HiFi-GAN unit-vocoder [13]. SeamlessM4T preserves elements of prosody and vocal style in all covered languages.

Machine translation. The MT landscape explores different approaches. Ott et al. [14] present NMT, a sequence-to-sequence machine translation Transformer model trained from scratch. The architecture takes a common vocabulary for each language pair, and the data are tokenized into sub-word units using the Byte-Pair Encoding (BPE) algorithm. Recent works investigate pre-training approaches. Liu et al. [15] introduce mBART, a sequence-to-sequence auto-encoder model pre-trained on a large-scale monolingual data in multiple languages with the BART objective [16]. This work emphasizes its advantage for languages not present in the pre-training data. Raffel et al. [17] propose T5, an encoder-decoder model with a transfer learning-based approach. Each textual data is treated as a text-to-text problem, enabling the model to perform multiple tasks (document summarization, machine translation, etc.) through a single model. The pre-training includes both supervised and unsupervised training on 20 TB of textual data from English, French, Romanian, and German languages. Finally, Costa-jussà [5] introduce NLLB, a massively multilingual Transformer-type model capable of automatically translating into 200 languages. This linguistic coverage can be beneficial between two related languages through interlinguistic transfer [18].

2.2. End-to-end approach

The second approach adapts end-to-end Speech Translation (ST) systems to our task. An ST model performs a direct translation from an audio sequence in a source language $s = (s_1, ..., s_{|s|})$ to the text $y = (y_1, ..., y_{|y|})$ in the target language. End-to-End ST circumvents the need for intermediary text and reduces the risk of error propagation.

In this work, we seek to leverage a system suitable for low-resource contexts. A recent work by Ye et al. [19] present ConST based on a contrastive learning approach. It aims to encode similar audio and textual representations in a close space. Comprising four modules, ConST integrates a vocal encoder using Wav2Vec 2.0 representations, a lexical embedding layer, and a Transformer encoder-decoder. The reported BLEU scores on MUST-C [20] demonstrate state-of-the-art performance, especially for low-resource language pairs.

[4] https://arxiv.org/abs/2312.05187

[5] https://arxiv.org/abs/2207.04672

3. Experiments

3.1. Dataset construction

We construct a dataset [6] from a pre-existing spontaneous speech corpora to train our approaches. Another dataset is built to evaluate this task. Each corpus C is a tuple (s, x, y) where $s = (s_1, ..., s_{|s|})$ is the audio segment, $x = (x_1, ..., x_{|x|})$ is the transcription, and $y = (y_1, ..., y_{|y|})$ is the pictogram terms translation.

Propicto-orféo is built upon the aligned speech/text data from the *Corpus d'Étude pour le Français Contemporain (CEFC)* [21]. We extracted 290,036 audio segments, representing 233 hours. From the transcriptions, we applied the method of Macaire et al. [22] to generate a pictogram-based translation, following specific rules and a restricted lexicon. Propicto-orféo covers multiple spontaneous speech situations, such as meetings and conversations in various domains.

We construct a dataset for evaluation, **Propicto-eval**, a corpus of read speech with 62 unique speakers, containing 3,011 sentences. The sentences are derived from children's stories, everyday situations, and sentences from the medical domain. These contexts are relevant as they mirror the types of interactions of our target audience. The dataset creation involved a three-step process: gathering sentences from publicly available ARASAAC PDFs, conducting a recording campaign spanning six months to get the corresponding audio, and generating pictogram translations using the method of Macaire et al. [22]. This process was overseen by the Data Protection Officer to ensure compliance with data protection rights.

3.2. Training details

Dataset pre-processing. We split the Propicto-orféo data into training, validation, and test sets, following an 80/10/10 distribution. We remove punctuation and convert transcriptions to lowercase. Each audio segment has a sampling rate of 16kHz with an upper duration limit of 30 seconds to maximize downstream performance. We select a subset of Propicto-eval comprising 100 sentences from 62 speakers with an equal distribution of female and male voices for evaluation. Data details are outlined in Table 2.

Table 2: *Distribution of data into three sets (train, development, test) of Propicto-orféo and Propicto-eval.*

	Propicto-orféo		**Propicto-eval**	
	# utterances	# hours	# utterances	# minutes
train	231,374	192	—	—
development	28,798	18	—	—
test	29,011	23	100	6

ASR training and inference. We use the SpeechBrain toolkit [23] and the provided recipe [7] to fine-tune the Wav2Vec2.0 model on Propicto-orféo, with a French pre-trained Wav2vec 2.0 model *LeBenchmark/wav2vec2-FR-7K-large* [24]. Audio segments of less than 3 seconds and longer than 10 seconds were excluded from the training to avoid empty audio segments and overfitting scenario (representing 45h). The training is performed with 4 Nvidia V100 GPUs with 32 GB of

[6] https://www.ortolang.fr/market/corpora/propicto

[7] https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech/ASR/CTC

memory each. We employ the latest released Whisper model namely *Whisper large-v3* [11] and follow the steps provided by the whisper repository [8] for inference. Finally, we use the HuggingFace Transformers library [25] to run the *SeamlessM4T-Large v2* model for ASR, setting the target language to French ('fra').

MT training. We use two toolkits for MT, Fairseq [26] and HuggingFace [25]. All experiments are performed on a single Nvidia V100 GPU with 32 GB of memory. We adapt the recipe provided by Fairseq to train the NMT model from scratch. A tokenization step with BPE segments the text into sub-word units. A vocabulary of 10,000 tokens is generated. The same toolkit is employed to fine-tune the *mbart-large-cc25* model learned from 25 languages. We adapt the suggested recipe[9] for translating from English into Romanian to our data. The *T5-large* and NLLB-200 (*facebook/nllb-200-1.3B*) models are fine-tuned by using the recipe[10] from HuggingFace for MT.

S2P training. We follow the pipeline integrated to Fairseq to train ConST[11]. The pre-trained French Wav2vec 2.0 model *LeBenchmark/wav2vec2-FR-7K-base* is initialized as the speech encoder, for computational reasons. A single Nvidia V100 GPU with 32 GB of memory is employed. The main parameters of each model[12] are described in Table 3.

Table 3: *Main parameters of ASR, MT and ST models for training and inference with the number of parameters, the learning rate, the batch size, the number of epochs and the total running time on Propicto-orféo training data.*

Model	# params	lr	# batch	# epochs	# time (h)
Whisper	1550M	—	—	—	—
SeamlessM4T	2.3B	—	—	—	—
Wav2Vec2-LeBenchmark	318,7M	1e-4	8	30	22.5
NMT	51M	5e-4	8	40	1.25
mBART25	610M	3e-5	8	40	18
T5-large	220M	2e-5	32	40	16
NLLB-200	600M	2e-5	32	40	30.5
ConST	150M	1e-4	8	40	100

4. Results and Discussion

4.1. ASR models

In the first set of experiments, we assess the performance of inference models in Table 4. Table 4 presents the Word Error Rate (WER) [27] on Propicto-orféo and Propicto-eval test sets. On Propicto-orféo, the WER with both the Whisper and Seamless models is recorded at 37.69 and 46.50, respectively. In contrast, the WER on Propicto-eval remains below 10%. This discrepancy suggests that inference models, having predominantly trained on read speech, therefore poorly generalize when applied to challenging corpora. In the second phase of our experiments, our objective is to contrast these results with a fine-tuned Wav2Vec2.0 approach on spontaneous speech (see Table 5). This approach, with a WER of 27.56 on the test set, demonstrates its effectiveness in handling spontaneous situations characterized by overlaps and disfluencies.

[8] https://github.com/openai/whisper

[9] https://github.com/facebookresearch/fairseq/tree/main/examples/mbart

[10] https://huggingface.co/docs/transformers/tasks/translation

[11] https://github.com/ReneeYe/ConST

[12] Please refer to the recipes quoted for full details of the parameters.

Table 4: *Word Error Rate (%) reported on Propicto-orféo and Propicto-eval test sets between two inference ASR models.*

Model	test – WER↓	
	Propicto-orféo	Propicto-eval
Whisper large-v3	37.69	9.01
SeamlessM4T-Large v2	46.50	**8.44**

Table 5: *Word Error Rate (%) reported on Propicto-orféo dev and test sets with a fine-tuned approach.*

Model	dev	test
Wav2Vec2.0-LeBenchmark	23.24	**27.56**

4.2. MT models

The results of the Machine Translation models are presented in Table 6. We report the BLEU score with sacreBLEU [28] on Propicto-orféo and Propicto-eval. The BLEU score was chosen because it offers greater nuance than the PER (Picto Error Rate) in a translation approach. The score is calculated by comparing the sequence of predicted pictogram terms with the sequence of source pictogram terms. For both corpora, mBART exhibits a substantial deviation from the other models, with a difference exceeding 12 BLEU points on Orfeo-picto. Moreover, NMT model trained from scratch outperforms mBART. When applying the models trained on Propicto-orféo to Propicto-eval, the results underscore the substantial contribution of multilingual pre-trained models T5 and NLLB to this translation task.

Table 6: *BLEU scores of the MT models on Propicto-orféo development and test sets and Propicto-eval test set.*

Model	dev	test	
	Propicto-orféo		Propicto-eval
NMT	**87.28**	**87.43**	69.89
mBART25	75.26	75.62	60.09
T5-large	85.21	85.88	73.51
NLLB-200	86.32	86.92	**79.25**

4.3. Combining ASR and MT for cascade S2P

We assess the performance of our cascade approach for Speech-to-Pictos translation by combining the ASR models with the MT systems. Table 7 presents the BLEU scores on the test data for each model combination with ASR inference models. The performance of Propicto-orféo experiences a significant decline when the translation system uses ASR system's predicted transcriptions as input to the MT model. In particular, we observe a decrease of over 28 points in the BLEU score, reaching 58.82 with the combination of Whisper and NLLB-200. The ASR system strongly influences the pictogram translation performance. On Propicto-eval, the best BLEU scores are given by NLLB-200 and T5-large with Whisper. We note a gap of 4.93 between the two MT models, which could be explained by the larger amount of data used to train NLLB, and therefore generalizes better to unseen data.

Finally, we compare the performance of the fine-tuned ASR approach with MT models on Propicto-orféo in Table 8. While

the NMT model performs the best in MT, NLLB-200 and T5-large with Wav2Vec2.0 achieve the highest BLEU scores in our cascade approach. We hypothesize that massively pre-trained multilingual models are more robust when dealing with terms distorted by the ASR system.

Table 7: *BLEU scores on Propicto-orféo and Propicto-eval test sets, with the combination of inference ASR models with the MT models trained on the Propicto-orféo training data.*

ASR model	MT model	test – BLEU↑	
		Propicto-orféo	Propicto-eval
Whisper large-v3	NMT	58.07	68.23
	mBART25	52.05	59.49
	T5-large	57.80	**72.25**
	NLLB-200	58.82	**77.18**
SeamlessM4T-Large v2	NMT	52.38	66.14
	mBART25	48.71	56.66
	T5-large	53.96	67.32
	NLLB-200	54.86	71.56

Table 8: *BLEU scores on Propicto-orféo test set by combining Wav2vec2.0 ASR model with the MT models.*

ASR model	MT model	Propicto-orféo
Wav2Vec 2.0-LeBenchmark	NMT	61.37
	mBART25	55.49
	T5-large	**61.66**
	NLLB-200	**62.48**

4.4. End-to-end S2P

We conclude our experiments by presenting the BLEU scores for the end-to-end speech translation model ConST on the Propicto-orféo development and test sets, as well as the Propicto-eval test set in Table 9. On clean data with a low Word Error Rate, the cascade approach outperforms ConST, as denoted by the BLEU score on Propicto-eval test set. However, in ecological acoustic conditions with Propicto-orféo, we observe nearly identical results with the cascade approach. Hence, we do not discount the potential of end-to-end approaches.

Table 9: *BLEU scores on Propicto-orféo dev and test sets with ConST, and on Propicto-eval test set, with the bracketed score showing the highest BLEU with the cascade approach.*

Model	dev	test	
	Propicto-orféo		Propicto-eval
ConST	62.21	60.21 (62.48)	54.47 (77.18)

4.5. Human evaluation

We perform a human evaluation, as the BLEU score does not offer precise insights into the specific behavior of each approach in the context of pictogram translation. This evaluation is performed by adapting an analytical framework MQM [29], which gives guidelines and procedures for measuring translation quality[13]. It determines whether the proposed translation meets the specifications agreed by the stakeholders. Each expert annotator assigns to each identified error in the text (source and/or target) a specific type and severity level. In this work, annotators had the option to select from 12 types of errors (addition, omission, unintelligible, etc.) and 4 severity levels (neutral, minor, major, critical).

In this study, 100 randomly chosen sentences from the Propicto-orféo test data were annotated by two expert annotators from the project, along with the sentences from Propicto-eval. Table 10 presents the Overall Quality Score (OQS) for the top two cascade models based on the highest BLEU score, and the end-to-end model ConST. OQS is calculated by multiplying the penalty score (resulting from the combination of the number of errors per category and the severity level, weighted accordingly) by the maximum value (usually 100). The translation system is not validated if the score is below the threshold value. Based on various observations, stakeholders have defined the limit for comprehension and usability of a translation to two major errors and one minor error, which corresponds to a threshold value of 89.

Table 10: *Overall quality score computed on 100 randomly annotated sentences of Propicto-orféo, and the test set of Propicto-eval of the two best cascade models and the end-to-end model.*

Model		OQS
Propicto-orféo	Wav2Vec2 + CTC / T5-large-orféo	45.78
	Wav2Vec2 + CTC / NLLB-200-orféo	44.56
	ConST-orféo	**62.62**
Propicto-eval	Whisper large-v3 / T5-large-orféo	75.29
	Whisper large-v3 / NLLB-200-orféo	**85.47**
	ConST-orféo	60.73

The best translation systems for Propicto-orféo and Propicto-eval fall below the set threshold, thereby rejecting their use with the target audience. This result can be attributed to certain behaviors, such as the inaccurate translation of named entities, or the mistranslation of specific terms (mainly due to errors introduced by the ASR systems). However, the OQS with ConST on Propicto-orféo stands at 62.62, exhibiting a gap of 16.84 compared to the best-performing cascade model. While automatic evaluation failed to validate the end-to-end approach, human evaluation underscores its efficacy in real-world acoustic scenarios. For Propicto-eval, the cascade approach demonstrates superior performance, with an OQS score nearly reaching the threshold value, which confirms its effectiveness to read speech in everyday situations. Future research could improve results by addressing untranslated and poorly translated terms, exploring novel end-to-end approaches, and employing generative systems for pictogram generation.

5. Conclusion

This article introduces two approaches for the automatic translation of speech into pictogram terms. We present data specifically designed for this task, encompassing various acoustic scenarios and domains. While the cascade approach demonstrates superior results compared to the end-to-end approach but competitive outcomes are observed with the latter on acoustically challenging data. Consequently, we do not discount the end-to-end approach for future exploration. Human evaluation exposes several limitations, including the substantial impact of speech recognition systems on translation and challenges in translating specific linguistic phenomena such as polylexical units and named entities.

[13] https://themqm.org/

6. Acknowledgements

This project was funded by the Agence Nationale de la Recherche (ANR) through the project PROPICTO (ANR-20-CE93-0005). This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011013625R1). We thank everyone who contributed to recording the sentences for the creation of the Propicto-eval corpus.

7. References

[1] D. R. Beukelman and P. Mirenda, *Communication alternative et améliorée: Aider les enfants et les adultes avec des difficultés de communication*. De Boeck Superieur, 2017.

[2] J. A. Pereira, D. Macêdo, C. Zanchettin, A. L. I. de Oliveira, and R. do Nascimento Fidalgo, "Pictobert: Transformers for next pictogram prediction," *Expert Systems with Applications*, vol. 202, p. 117231, 2022.

[3] E. Cataix-Nègre, *Communiquer autrement: Accompagner les personnes avec des troubles de la parole ou du langage*. De Boeck Superieur, 2017.

[4] C. Vaschalde, P. Trial, E. Esperança-Rodier, D. Schwab, and B. Lecouteux, "Automatic pictogram generation from speech to help the implementation of a mediated communication," in *Conference on Barrier-free Communication*, 2018.

[5] V. Vandeghinste, I. S. L. Sevens, and F. Van Eynde, "Translating text into pictographs," *Natural Language Engineering*, vol. 23, no. 2, pp. 217–244, 2017.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[12] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 244–250.

[13] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[14] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling neural machine translation," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1–9.

[15] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2020, pp. 8440–8451.

[19] R. Ye, M. Wang, and L. Li, "Cross-modal contrastive learning for speech translation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5099–5113.

[20] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2012–2017.

[21] C. Benzitoun, J.-M. Debaisieux, and H.-J. Deulofeu, "Le projet orféo: un corpus d'étude pour le français contemporain," *Corpus*, no. 15, 2016.

[22] C. Macaire *et al.*, "A multimodal French corpus of aligned speech, text, and pictogram sequences for speech-to-pictogram machine translation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, May 2024, pp. 839–849.

[23] M. Ravanelli *et al.*, "SpeechBrain: A general-purpose speech toolkit," 2021.

[24] S. Evain *et al.*, "LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech," in *INTERSPEECH 2021: Conference of the International Speech Communication Association*, Brno, Czech Republic, Aug. 2021.

[25] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[26] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[27] J. Woodard and J. Nelson, "An information theoretic measure of speech recognition performance," *Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA*, 1982.

[28] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191.

[29] A. Burchardt, "Multidimensional quality metrics: a flexible system for assessing translation quality," in *Proceedings of Translating and the Computer 35*. London, UK: Aslib, Nov. 28-29 2013.