A Comprehensive Analysis of Tokenization and Self-Supervised Learning in End-to-End Automatic Speech Recognition applied on French Language

Thibault Bañeras-Roux, Mickael Rouvier, Jane Wottawa, Richard Dufour

A Comprehensive Analysis of Tokenization and Self-Supervised Learning in End-to-End Automatic Speech Recognition applied on French Language

Thibault Bañeras-Roux
Nantes University
LS2N
Nantes, France

Mickael Rouvier
Avignon University
LIA
Avignon, France

Jane Wottawa
Le Mans University
LIUM
Le Mans, France

Richard Dufour
Nantes University
LS2N
Nantes, France

***Abstract*—The performance of end-to-end automatic speech recognition (ASR) systems enables their increasing integration into numerous applications. While there are various benefits to such speech-to-text systems, the choice of hyperparameters and models plays a crucial role in their performance. Typically, these choices are determined by considering only the character (CER) and/or word error rate (WER) metrics. However, it has been shown in several studies that these metrics are largely incomplete and fail to adequately describe the downstream application of automatic transcripts. In this paper, we conduct a qualitative study on the French language that investigates the impact of subword tokenization algorithms and self-supervised learning models from different linguistic and acoustic perspectives, using a comprehensive set of evaluation metrics.**

***Index Terms*—automatic speech recognition, evaluation metrics, tokenization, self-supervised learning**

I. Introduction

Automatic Speech Recognition (ASR) technology is integral to various applications, including transcription services, voice assistants, and automated captioning. Its ability to convert spoken language into written text has significantly enhanced the accessibility and usability of audio content. With the ever-increasing demand for precise and efficient ASR systems, researchers are continuously exploring innovative methods to enhance their performance.

ASR models heavily rely on tokenization as a foundational element in the transcription process. Traditionally, word tokenization segments text into individual words using predefined delimiters like spaces and punctuation marks. ASR systems predict these tokens with the assistance of a decoder. Modern ASR systems employ a more sophisticated tokenization approach, segmenting words into smaller units known as subwords. This finer tokenization enhances the system's ability to handle out-of-vocabulary (OOV) words and reduces vocabulary size. Among the prominent tokenization approaches in use, we can mention Byte-Pair Encoding (BPE) [1] or SentencePiece [2].

Another critical aspect shaping the advancement of ASR systems is Self-Supervised Learning (SSL). Notably, the development of wav2vec [3] and [4] has significantly bolstered the acoustic generalization capabilities of ASR systems. These SSL models are trained without the need for manual annotations, leveraging vast amounts of audio data to generate speech representations known as *embeddings*. These embeddings serve as concise representations of speech segments, capturing essential acoustic characteristics from the speech data. When integrated into ASR systems, they substantially enhance adaptability to various speaking styles, accents, and background noise, resulting in more robust and accurate speech recognition.

However, the impact of these parameters remains relatively unexplored within ASR research, particularly on end-to-end ASR systems, eliminating the need for intermediate representations or separate processing stages. While these architectures are gaining importance, our understanding of them is still in its early stages. Evaluation often focuses predominantly on metrics such as Word Error Rate (WER), while broader implications of tokenization and SSL choices on transcription quality are seldom examined or rigorously investigated. This paper aims to fill this gap. Building upon previous work [5], we propose a comprehensive study examining the effects of tokenizers and SSL models on lexical, acoustic, and semantic metrics [6]–[8] specifically tailored to the French language.

In this paper, we make the following contributions:

- We establish that a reduced vocabulary enhances the generalization capabilities of ASR systems.
- We provide compelling evidence for the effectiveness of using the Unigram tokenizer with a reduced vocabulary, particularly in the context of French language ASR, resulting in improved performance.
- We demonstrate that evaluation criteria for metrics have a discernible impact at the system level. Consequently, a system deemed optimal by WER may not necessarily be the best from a human, semantic, or other perspectives.
- Our results suggest that the use of phonetically adapted tokens does not yield performance improvements compared to traditional tokenization methods.

The paper is organized as follows. In Section II, we describe the methodology employed for evaluating the different tokenizers and SSL models, and discuss the evaluation metrics.

Then, we study how the choice of self-supervised models influences ASR performance in Section III and we carry out an analysis of the impact of tokenizer hyperparameters in Section IV. In Section V, we discuss the difficulties in clearly assessing the ASR performance, with the nature of the evaluation metric influencing the reported quality of a transcription system. We finally conclude in Section VI.

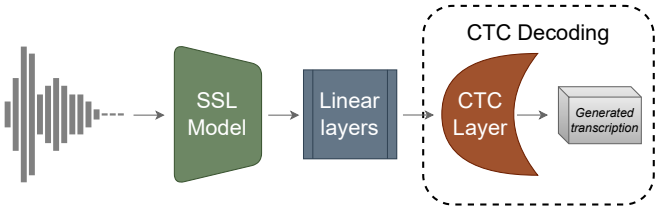

Fig. 1: Architecture of the Automatic Speech Recognition systems used in this study.

II. Study Methodology

In this section, we detail our study setup. Firstly, we discuss two key components, including their studied approaches: tokenization strategies (Section II-A) and self-supervised learning models (Section II-B). Then, we include a description of corpora (Section II-C) and evaluation metrics (Section II-E) to study various performance aspects. Finally, we describe the used end-to-end ASR system (Section II-D).

A. Tokenization strategies

Tokenization techniques have undergone significant evolution, employing diverse algorithms to process language. Tokenization involves breaking raw text into smaller units known as tokens, which can represent words, subwords, or characters. In end-to-end ASR, the prevailing approach computes a sequence of token probabilities for individual speech segments, generating transcriptions using the Connectionist Temporal Classification (CTC) framework.

Among the most used tokenization strategies, Byte-Pair Encoding (BPE) tokenization [1] segments text into subwords using a vocabulary initialized with characters and expanded through an iterative merging of the most frequent token pairs. WordPiece [9] shares similarities with BPE but adopts a likelihood-based merging approach. Unigram [10], on the other hand, focuses on trimming a large vocabulary using loss-based criteria. Additionally, SentencePiece [10] implements both Unigram and BPE but does not pre-tokenize sentences into words. These subword tokenization strategies are summarized in Table I.

We employ various tokenization methods, including character, BPE, SentencePiece, and Unigram, while varying the vocabulary size for subword tokenization. Given that speech is the primary modality in ASR, investigating the use of grapheme-based linguistic tokenization may be pertinent. Unlike tokens chosen based on the co-occurrence frequency of characters and subwords, grapheme-based tokenization fully respects linguistic and acoustic characteristics. To this end, we train systems with a vocabulary based on 144 graphemes[1] (character string transcribing a phoneme), compiled by cross-referencing various teaching resources. With BPE, the vocabulary can be initialized with a specific set of tokens, not just characters. Thus, we initialize BPE tokenizers with graphemes.

B. Self-supervised learning (SSL) models

We employed multiple SSL models, each trained on diverse datasets and languages. Specifically, we used LeBenchmark large models [11], which are wav2vec 2.0 models trained on different amounts of French data: 1,000 hours (w2v2-FR-1k), 3,000 hours (w2v2-FR-3k), and 7,000 hours (w2v2-FR-7k). This allowed us to assess ASR performance across various amounts of training data. Additionally, we included the classic wav2vec 2.0 model (w2v2-EN-53k) [3], trained on 53,000 hours of English data, to investigate its transfer-learning capabilities to another language. Furthermore, we incorporated an XLSR model (w2v2-xlsr) [12], which is a wav2vec 2.0 model trained on a diverse dataset comprising 53 languages, including French. This enabled us to examine the effects of cross-lingual training on ASR performance. Using these SSL models, our goal was to gain a comprehensive understanding of the impact of training data size and monolingual/multilingual training, focusing on identifying the most effective approaches for developing end-to-end ASR systems.

C. Corpora

The relationship between spoken and written French is intriguing due to the relatively large presence of silent letters, which can induce distinctive behavior of semantic and lexical metrics at the word and character level. Therefore, all end-to-end ASR systems have been trained to process French using ESTER 1 [13] and ESTER 2 [14], EPAC [15], ETAPE [16] and REPERE [17] train corpora. Collectively, these corpora represent approximately 356 hours of audio, comprised of radio and television broadcast data.

Our comprehensive analysis is based on the French REPERE test corpus, which corresponds to 10 hours of speech.

D. Automatic Speech Recognition systems

In this study, we set up 28 end-to-end ASR systems based on the Speechbrain toolkit [18]. All the ASR systems incorporate an SSL model, a Deep Neural Network (DNN) layer composed of three linear layers and a CTC layer, as shown in Figure 1. These systems are trained for 10 epochs using CTC loss on the corpora described in Section II-C with a lower learning rate for the SSL model. For inference, the transcription is generated with *best path decoding*. For reproducibility, settings are detailed in our GitHub code repository[2].

[1] aa, ae, aen, ai, aï, ail, aim, ain, am, an, aon, aou, au, aw, ay, aye, bb, ca, cc, cca, cce, cch, cci, cco, ccu, ccueil, ccy, ce, ch, ci, co, cqu, ct, cu, cueil, cy, dd, ds, ea, ean, eau, ect, ed, ee, ée, ef, ei, eil, eim, ein, em, emmm, en, enn, ent, er, es, eu, eû, ew, ez, ff, ga, ge, geu, geü, gg, gge, ggi, gh, gi, gn, go, gt, gu, gua, gue, guë, güe, gui, ign, iil, il, ill, illaire, ille, illier, im, imm, imma, imme, immi, immo, immu, in, ing, ll, lle, mm, mn, nn, oa, oe, oi, oil, om, on, ou, ph, pp, ps, pt, qu, qua, qui, rh, rr, rrh, sc, sca, sce, sch, sci, sco, scu, scy, ss, th, tia, tie, tiel, tien, tient, tieuse, tieux, tion, tt, tz, uil, um, un, uy, ym, yn

[2] https://github.com/thibault-roux/systems-analysis

Tokenizer	Description	Word split	Operation	Scoring
BPE	Initializes the vocabulary with individual characters and iteratively merges the most frequent token pairs until the desired vocabulary size is achieved.	Yes	Merge	Frequency
WordPiece	Similar to BPE, it selects the token pair that maximizes the likelihood of the training data, rather than choosing the most frequent pair.	Yes	Merge	Likelihood
Unigram	Initializes the vocabulary with a large number of tokens and then systematically reduces the size of the vocabulary by iteratively trimming each token.	No	Trim	Likelihood
SentencePiece	Tool that employs BPE or Unigram algorithms without segmenting the training data into words.	No	Both	Both

TABLE I: Overview of the most commonly used subword tokenizers.

SSL model	WER	CER	SemDist	UWER	PhonER
w2v2-FR-1K	18.94	7.63	12.52	77.42	6.26
w2v2-FR-3K	17.16	6.87	11.20	76.84	5.44
w2v2-FR-7k	**16.56**	**6.72**	**10.45**	**75.19**	**5.29**
w2v2-xlsr	21.48	8.59	14.47	78.66	7.03
w2v2-EN-53k	36.41	13.67	23.62	89.83	12.63

TABLE II: Performance of ASR systems using a character tokenizer and different SSL models (French models with w2v2-FR-1k, w2v2-FR-3k, and w2v2-FR-7k; English model with w2v2-EN-53k; multilingual model with w2v2-xlsr).

E. Evaluation metrics

Instead of focusing solely on the classical WER metric, we examine various aspects of automatic transcriptions using metrics that evaluate lexical, semantic, and acoustic levels.

For the lexical aspect, we consider classical metrics such as **Word Error Rate (WER)** and **Character Error Rate (CER)**. Additionally, inspired by the Individual Word Error Rate [19] and aiming to study the generalization ability of ASR systems, we developed the **Unseen Word Error Rate (UWER)**. The UWER measures the accuracy of transcribed words specifically for those absent from the training corpora but present in the test set, providing a valuable assessment of the system's ability to generalize to unseen vocabulary.

At the semantic level, we employ the **SemDist** [7] metric, which computes the cosine similarity between embeddings of the reference and hypothesis obtained at the sentence level. In our experiments, we utilized a sentence embedding model[3] (SentenceBERT [20]) based on CamemBERT [21], a French pre-trained BERT version. This metric had the strongest correlation with human perception in a previous study [22].

In addition to text transcripts derived from speech, we also consider an acoustic metric: the **Phoneme Error Rate (PhonER)**, which involves computing the Levenshtein distance between reference and hypothesis sequences of phonemes both obtained using an automatic grapheme-to-phoneme converter[4].

III. Impact of SSL models

In this section, we reproduce previous results [23] on how the language (Section III-A) and the size (Section III-B) of the training data used by SSL models affect the end-to-end ASR system's performance and deepen this analysis by using several metrics. To ensure a fair comparison between SSL models, all the results in this section use a character tokenizer. Table II presents the performance obtained by our end-to-end ASR system using various SSL model configurations.

A. Impact of training language

As depicted in Table II, SSL models pre-trained on French data (*w2v2-FR-**) consistently demonstrate superior performance across metrics. Conversely, the English-based system (*w2v2-EN-53k*), despite having the largest training dataset, exhibits a relatively high Word Error Rate (WER) of 36.41%. In contrast, the ASR system trained on the target language achieves a substantially improved WER of 16.52% using the same character tokenizer (*w2v2-FR-7k*). This highlights the significance of training SSL models on the target language to acquire language-specific knowledge crucial for accurate transcription.

When fine-tuning an SSL model trained on a diverse dataset that includes the target language (*w2v2-xslr*), we observe a performance drop compared to the monolingual French system, resulting in a WER of 21.48%. However, this multilingual system still outperforms the English-based ASR system. It is important to note that in multilingual training, there is a risk that language-specific information may become overwritten, diluted, or averaged by the inclusion of other languages.

B. Impact of training data size

We now narrow our focus to the analysis of the French models only (*w2v2-FR-**), as presented in Table II, to examine the impact of SSL training data size. Our analysis reveals a clear and direct correlation between the size of the training data and improved performance across all considered metrics. Increasing the training data size enables the model to learn more comprehensive representations and better capture a wider range of acoustic and linguistic variations.

IV. Impact of Tokenization Strategies

In this section, we explore how tokenization algorithms can impact the performance assessment of ASR systems. We begin by comparing subword units and character tokenization (Section IV-A). Then, we assess the use of graphemes as

[3]https://huggingface.co/dangvantuan/sentence-camembert-large

[4]https://github.com/Remiphilius/PoemesProfonds

Tokenizer	# Token	WER	CER	SemDist	UWER	PhonER	Avg. token
BPE	1000	15.98	7.00	10.08	78.74	5.72	1.92
	750	15.33	6.67	9.41	76.67	5.31	2.05
	500	15.57	6.73	9.61	76.43	5.38	2.28
	250	15.16	6.45	9.43	74.11	5.05	2.75
	150	15.47	6.46	9.47	74.77	5.10	3.20
BPE with graphemes	1000	15.74	6.62	9.97	77.25	5.40	2.76
	750	15.98	6.63	10.03	77.58	5.47	2.81
	500	15.64	6.59	9.77	76.51	5.34	2.93
	250	15.74	6.55	9.73	75.77	5.18	3.10
SentencePiece	1000	15.78	6.87	9.76	77.83	5.14	1.88
	750	15.59	6.76	9.39	76.18	5.35	2.03
	500	15.51	6.66	9.55	76.43	5.33	2.26
	250	15.70	6.74	9.75	74.52	5.37	2.75
	150	15.56	6.57	9.52	74.52	5.58	3.29
Unigram	1000	15.49	6.68	9.57	78.91	5.37	1.88
	750	15.29	6.55	9.34	76.67	5.23	2.03
	500	15.54	6.70	9.57	76.26	5.29	2.26
	250	15.58	6.65	9.44	73.53	5.23	2.77
	150	**15.07**	**6.36**	**9.33**	**73.12**	**4.90**	3.33
Character	-	16.56	6.72	10.45	75.19	5.29	4.88

TABLE III: Performance of ASR systems using different tokenizers (BPE, character, graphemes, SentencePiece and Unigram).

subword units to determine the most suitable subword unit for optimizing ASR system performance (Section IV-B).

Table III presents the performance of end-to-end ASR systems trained with different tokenization strategies. The last column of the table, *Avg. token*, represents the average subword units per word for each tokenizer on the test dataset. To ensure a fair comparison between tokenization strategies for all ASR systems, we employed the SSL model *w2v2-FR-7k*, known for its superior performance. Notably, the system utilizing the Unigram tokenizer with a fixed vocabulary of 150 consistently achieved the best results across various metrics, including WER, CER, SemDist, UWER, and PhonER.

A. Subword units vs Character tokenization

We observe in Table III that subword unit tokenizers (BPE, SentencePiece, and Unigram) consistently outperform character tokenization since this tokenization neglects linguistic and acoustic intricacies in speech. In contrast, subword unit tokenizers capture more nuanced and contextually relevant information, which explains their better performance across all metrics.

B. Influence of graphemes

It is noteworthy that the *BPE with grapheme* tokenizer consistently yields inferior results compared to other subword unit tokenizers. Despite its intention to integrate knowledge about acoustics and linguistics, end-to-end ASR models struggle to effectively utilize this information, resulting in suboptimal outcomes.

Table IV displays the percentage of graphemes included in the vocabulary of other tokenizers. An interesting finding is that the best-performing system has the lowest percentage of graphemes. This observation, coupled with the lower relative performance of systems using graphemes, suggests that subword units align more closely with linguistic elements than with acoustics.

# Token	SentencePiece	BPE	Unigram
250	17.24%	17.24%	11.03%
500	21.38%	21.38%	17.93%
750	24.14%	24.14%	22.07%
1000	27.59%	28.97%	23.45%

TABLE IV: Percentages of graphemes included in the vocabulary of different tokenizers.

V. Metrics discrepancy

Despite the 150-size Unigram tokenizer outperforming others for all metrics (as demonstrated in Table III), metrics fail to establish a consistent ranking between systems. For instance, for SentencePiece, the best system has a vocabulary size of 500 according to WER, 150 according to CER and UWER, and PhonER, and 750 according to SemDist. Table V illustrates the Spearman correlation between metrics at the system level, revealing that the indicated hierarchy can vary significantly.

The discrepancies between metrics pose challenges in determining a clear best system because different metrics offer conflicting rankings. This inconsistency prompts questions about the relevance of standard metrics like WER for accurately evaluating system performance. Previous research [22] has already shown that metrics do not equally correlate with human perception. In the context of French and across a range of metrics assessing aspects like lexical accuracy, semantics, and phonetics, it was observed that, at the utterance level, WER had one of the lowest correlations with human perception, while SemDist, using sentence embeddings, exhibited the strongest correlation. In our study, these differences underscore that ASR metrics can yield varying assessments of performance at the system level, which is a first, to our knowledge.

VI. Conclusion and perspectives

In this paper, we conducted a thorough analysis of two pivotal factors influencing ASR system performance: tokenization strategy and self-supervised learning (SSL) models. Our

	WER	CER	SemDist	UWER	PhonER
WER					
CER	0.55				
SemDist	0.87	0.45			
UWER	0.34	0.45	0.47		
PhonER	0.63	0.76	0.61	0.80	

TABLE V: Spearman correlation of metrics at system level.

findings shed light on the intricate relationship between these components and various language aspects, offering valuable insights for the speech community.

Regarding tokenization, our analysis unveiled that systems with larger vocabulary sizes encountered challenges in generalizing to out-of-vocabulary (OOV) words. Conversely, character tokenization excelled in terms of Character Error Rate (CER) but faced difficulties in maintaining lexical accuracy and word boundaries.

In the realm of SSL models, we corroborate the conclusions of previous works [23] by observing a direct correlation between training data size and improved ASR system performance across all metrics. Larger SSL model training datasets in the target language facilitate better generalization and enhanced representation learning, resulting in overall improved performance. Additionally, our study underscores the significance of pre-training SSL models on the target language, as models not specifically trained on it exhibited performance limitations due to the lack of language-specific knowledge.

A significant outcome of our study is the inconsistency among evaluation metrics in determining a clear best-performing ASR system. While various metrics have been employed, they exhibited divergent rankings, challenging their ability to comprehensively assess system performance. These discrepancies underscore the necessity to explore alternative evaluation approaches for both intrinsic and downstream evaluations tailored to the task at hand.

ACKNOWLEDGMENT

This work was financially supported by the DIETS project financed by the Agence Nationale de la Recherche (ANR) under contract ANR-20-CE23-0005.

REFERENCES

[1] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), 2016.

[2] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018.

[3] 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*, 2020.

[4] 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*, 2021.

[5] S. Singh, A. Gupta, A. Maghan, D. Gowda, S. Singh, and C. Kim, "Comparative study of different tokenization strategies for streaming end-to-end asr," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021.

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[7] S. Kim, A. Arora, D. Le, C.-F. Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding," in *Interspeech*, 2021.

[8] T. Bañeras-Roux, M. Rouvier, J. Wottawa, and R. Dufour, "Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition," in *Interspeech 2022*, 2022.

[9] M. Schuster and K. Nakajima, "Japanese and korean voice search," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012.

[10] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[11] S. Evain, M. H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet *et al.*, "Task agnostic and task specific self-supervised learning from speech with lebenchmark," in *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

[12] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," 2021.

[13] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News," in *International Conference on Language Resources and Evaluation (LREC)*, 2006.

[14] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[15] Y. Esteve, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas, "The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news," in *International Conference on Language Resources and Evaluation (LREC)*, 2010.

[16] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *International Conference on Language Resources and Evaluation (LREC)*, 2012.

[17] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition," in *International Conference on Language Resources and Evaluation (LREC)*, 2012.

[18] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh *et al.*, "SpeechBrain: A general-purpose speech toolkit," 2021.

[19] S. Mdhaffar, Y. Estève, N. Hernandez, A. Laurent, R. Dufour, and S. Quiniou, "Qualitative evaluation of asr adaptation in a lecture context: Application to the pastel corpus." in *Interspeech*, 2019.

[20] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[21] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. De La Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a Tasty French Language Model," in *58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[22] T. Bañeras-Roux, J. Wottawa, M. Rouvier, T. Merlin, and R. Dufour, "HATS: An Open data set Integrating Human Perception Applied to the Evaluation of Automatic Speech Recognition Metrics," in *Text, Speech and Dialogue*, 2023.

[23] S. Evain, H. Nguyen, H. Le, M. Zanon Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, "LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech," in *INTERSPEECH 2021: Conference of the International Speech Communication Association*, 2021.