



**HAL**  
open science

# HATS: An Open data set Integrating Human Perception Applied to the Evaluation of Automatic Speech Recognition Metrics

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

## ► To cite this version:

Thibault Bañeras-Roux, Jane Wottawa, Mickael Rouvier, Teva Merlin, Richard Dufour. HATS: An Open data set Integrating Human Perception Applied to the Evaluation of Automatic Speech Recognition Metrics. Text, Speech and Dialogue 2023 - Interspeech Satellite, Faculty of Applied Sciences University of West Bohemia Plzeň (Pilsen); NTIS P2 Research Center University of West Bohemia Plzeň (Pilsen); Faculty of Informatics Masaryk University Brno, Sep 2023, Plzeň (Pilsen), Czech Republic. hal-04125590

**HAL Id: hal-04125590**

**<https://hal.science/hal-04125590>**

Submitted on 12 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# HATS: An Open data set Integrating Human Perception Applied to the Evaluation of Automatic Speech Recognition Metrics

Thibault Bañeras-Roux<sup>1</sup>, Jane Wottawa<sup>2</sup>, Mickael Rouvier<sup>3</sup>, Teva Merlin<sup>3</sup>, and Richard Dufour<sup>1</sup>

<sup>1</sup> Nantes University, LS2N

<sup>2</sup> Le Mans University, LIUM

<sup>3</sup> Avignon University, LIA

{thibault.roux, richard.dufour}@univ-nantes.fr

{mickael.rouvier, teva.merlin}@univ-avignon.fr

jane.wottawa@univ-lemans.fr

**Abstract.** Conventionally, Automatic Speech Recognition (ASR) systems are evaluated on their ability to correctly recognize each word contained in a speech signal. In this context, the word error rate (WER) metric is the reference for evaluating speech transcripts. Several studies have shown that this measure is too limited to correctly evaluate an ASR system, which has led to the proposal of other variants of metrics (weighted WER, BERTscore, semantic distance, etc.). However, they remain system-oriented, even when transcripts are intended for humans. In this paper, we firstly present **Human Assessed Transcription Side-by-side (HATS)**, an original French manually annotated data set in terms of human perception of transcription errors produced by various ASR systems. 143 humans were asked to choose the best automatic transcription out of two hypotheses. We investigated the relationship between human preferences and various ASR evaluation metrics, including lexical and embedding-based ones, the latter being those that correlate supposedly the most with human perception.

**Keywords:** automatic speech recognition, evaluation metrics, human perception, manual annotation

## 1 Introduction

Automatic Speech Recognition (ASR) consists in transcribing speech into its textual form. Automatic transcriptions can for example be used by humans in the case of captioning, speech-to-text messages or by third systems such as virtual personal assistants. Since the emergence of hidden Markov model-based ASR systems [18] for processing continuous speech, the field has seen an important breakthrough with the use of deep neural networks and self-supervised methods such as wav2vec [1] and HuBERT [16]. These approaches allow the extraction of meaningful information from speech without previously labeled data.

Faced with transcription errors, unlike a machine, a human is able to process the sentence anyway and extract its initial meaning if the latter was not fundamentally

impacted by the errors. Errors in automatic transcriptions can arise due to various factors such as noise in the speech signal, speaker accents, or technical limitations. The question is to determine which errors are acceptable and which ones may cause comprehension difficulties for humans. Thus, it is crucial to evaluate the quality of automatic transcriptions based on their overall comprehensibility to humans.

Currently, the most commonly used metrics for evaluating ASR systems are the Word Error Rate (WER), which measures the number of incorrectly transcribed words, and the Character Error Rate (CER), which calculates the number of characters that differ from the reference transcription. However, many researchers [33,7,17,19] have pointed out issues with these metrics, such as the absence of error weighting or the lack of linguistic and semantic knowledge. Consequently, there has been a growing interest in developing new metrics to evaluate ASR systems. Some researchers [23,27,13,20,2] have therefore started exploring alternative metrics that can more accurately assess the quality and effectiveness of automatic transcriptions. Similarly, these issues have been observed in the field of machine translation. As a result, new metrics and data sets have been produced from multiple shared tasks [26,25,9,8]. Semantic-based metrics, such as BERTScore [34], have then been shown to be effective in evaluating the quality of machine-generated translations.

While these metrics are obtained automatically and are rather *machine-oriented*, human evaluations of ASR systems have been carried out in the past, which includes side-by-side experiments [13,19,21] where human subjects are asked to choose the best transcript among two options. These studies have also enabled assessing the quality of automatic metrics from a human perspective. The present study builds on these side-by-side experimental protocols, but instead of modifying the speech signal or text hypothesis with artificially generated errors, or using different outputs from the same ASR system to obtain two different hypotheses, our study utilizes the outputs of ten ASR systems with varying architectures applied on the same speech corpus. Furthermore, rigorous criteria were used to select the transcripts where choices are the harder in order to study metric and human behavior. The advantage of the side-by-side experiment is that the subject has to make a choice between two hypotheses, which does not allow for equality. In contrast to direct assessment, side-by-side experiments eliminate the potential bias of prior choices, allowing for consistent comparisons between transcriptions. By comparing human judgments to those of the metric, we can effectively evaluate its performance.

In this paper, we introduce HATS (Human-Assessed Transcription Side-by-Side), a new open data set of human preferences on erroneous transcriptions in French from various ASR architectures. As a second contribution, an original study is conducted using HATS to evaluate automatic metrics by analyzing their agreement with human assessments. Our objective is to identify the ASR evaluation metrics that most closely correlate with human perception. The HATS data set is freely released to the scientific community<sup>4</sup>.

The paper is organized as follows: Section 2 describes the used ASR systems and the automatic metrics that will be evaluated based on their correlation with human perception. In Section 3, we present the implementation of the side-by-side human perception experiment, including the protocol for selecting the transcripts provided to

<sup>4</sup> <https://github.com/thibault-roux/metric-evaluator>

human evaluators. Section 4 describes the HATS data set, while Section 5 presents a study on the quality of automatic metrics for evaluating transcription systems in relation to human perception. Finally, Section 6 provides the conclusion and future work.

## 2 Transcription systems and ASR evaluation metrics

In Section 2.1, we present the different automatic speech recognition systems used to obtain the automatic transcriptions that constitute the HATS corpus. Then, in Section 2.2, we describe all the evaluation metrics applied to assess these transcriptions and evaluate them in relation to human perception.

### 2.1 Automatic transcription systems

In this study, we set up 8 end-to-end systems based on the Speechbrain toolkit [30] and 2 DNN-HMM-based systems using a state-of-the-art recipe<sup>5</sup> with the Kaldi toolkit [29]. The end-to-end ASR systems were trained using various self-supervised acoustic models. Seven of the systems used variants of the wav2vec2 models learned on French [6], and one system used the XLS-R-300m model. In the Kaldi pipeline systems, one of the systems included an extra rescoring step using a neural language model.

All ASR systems have been trained to process French using ESTER 1 and 2 [10,11], EPAC [5], ETAPE [15], REPERE [12] train corpora, as well as internal data. Taken together, the corpora represent approximately 940 hours of audio comprised of radio and television broadcast data. The transcripts used to build our HATS corpus are extracted from the REPERE test set, which represents about 10 hours of audio data.

### 2.2 Evaluation metrics

We propose to focus on evaluation metrics for transcription systems that enable us to evaluate the systems at both lexical and semantic levels. First of all, we consider classical lexical metrics such as **Word Error Rate** and **Character Error Rate**.

Next, we examine three semantic metrics based on word embedding representations. The first one, **Embedding Error Rate (EmBER)** [2], is a WER where substitution errors are weighted according to the cosine distance between the reference and the substitute word embeddings obtained from fastText [14,3]. The second one, **SemDist** [20], involves calculating the cosine similarity between the reference and hypothesis using embeddings obtained at the sentence level. We compared different pre-trained word embedding models to evaluate their impact on the metric. Specifically, we compared using the embedding of the first token from CamemBERT [24] or FlauBERT [22] models, or using the output of a sentence embedding model (SentenceBERT [31]). Our last semantic metric is **BERTScore** [34], that computes a similarity score for each token in the candidate sentence with each token in the reference sentence using contextual embeddings. In our study, we use a multilingual BERT [4] and CamemBERT<sup>6</sup> [24] models (both CamemBERT-base and CamemBERT-large).

<sup>5</sup> <https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/>

<sup>6</sup> <https://camembert-model.fr>

While text transcriptions are derived from speech, we also consider a **Phoneme Error Rate (PER)**, which involves computing the Levenshtein distance between reference and hypothesis sequences of phonemes obtained using a text-to-phoneme converter<sup>7</sup>.

### 3 Side-by-side human evaluation protocol

This section describes the collection of the HATS corpus. The setup of the perceptual experiment is summarized in Section 3.1, while the protocol for selecting automatic transcripts for human evaluation is described in Section 3.2.

#### 3.1 Perceptual experiment

In our study, the side-by-side experiment involves presenting the subject with a manually transcribed reference to represent the speech, as well as two automatic transcripts, each produced by a different system. The automatic transcriptions always contained errors with respect to the reference. Each triplet comprised of a reference and two hypotheses is called a stimulus to which participants react in choosing their preferred hypothesis. In the following, *stimuli* refers to the different triplets to which each participant was confronted.

The experiment was made available online which allowed for participants to realize the task remotely and at their preferred time. They used a mouse to choose their preferred hypothesis according to the reference. The study utilized a minimal instruction protocol (See Figure 1), which allowed participants to self-determine the criteria that were important in determining the quality of a transcript. Figure 1 illustrates the visual display presented to the subjects during the study. The reference was in written form only, in order to allow a comparison of ASR-oriented metrics and human perception within the same context [32].


To avoid possible biases, the stimuli were presented in a random order, both for the order of the triplet, and for the order of the two hypotheses (the same hypothesis can be A or B).

For this study, 143 online participants volunteered. Before starting with the evaluation, they filled out a questionnaire helping to assess their age, spoken languages, and level of education. All participants are fluent in French and have an average age of 34 years with a standard deviation of 13.5 years. In Figure 2 and Figure 3, we can see the distribution of number of spoken languages and education level for our studied population. Each participant evaluated 50 triplets of transcripts in random order, for a total time of about 15 minutes per participant.

#### 3.2 Protocol for stimuli selection

The transcription triplets coming from the REPERE test corpus were not selected randomly. In this study, we attach great importance to the selection of stimuli and we decided to study human behavior and metrics in complex situation, i.e. where humans

<sup>7</sup> <https://github.com/Remiphilius/PoemesProfonds>

Progress bar : 

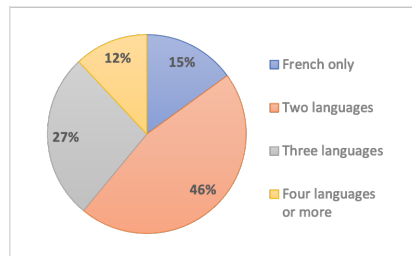
We call reference an exact transcription from audio to text.  
We propose two hypotheses (called transcription) produced by speech recognition systems. Choose the transcription that seems to you the most acceptable.

Reference :

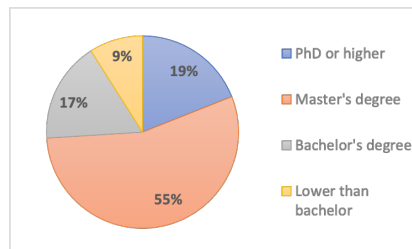
**Transcription 1 :**  
how are you two day patrick

**Transcription 2 :**  
were you today patrick

**Fig. 1.** Screenshot from the side-by-side experiment.



**Fig. 2.** Participant characterization in terms of number of spoken languages.



**Fig. 3.** Participant characterization in terms of level of education.

have difficulties to choose the best transcription. In this context, the aim was to maximize the diversity of choices to be made: subjects had to choose among errors made by different systems (since it is unlikely that different systems produce identical errors). Also, it would be interesting to study the cases where the choice is easy for the automatic metrics, as well where ambiguous scores are obtained, or where two metrics disagree to determine which one of the two hypotheses is best.

Therefore, the following three criteria had to be respected: (1) both hypotheses must be different from each other and have at least one character that differs from the reference, (2) hypotheses from every system were contrasted with hypotheses from every other system, and (3) hypotheses pair selection was based on metric scores. The selection criteria (3) based on the metrics can be divided into three different categories: (A) each metric was compared to itself presenting either the same, a slightly different or a highly different score between the two hypotheses, (B) in both hypotheses the WER or CER were equal but WER or CER, EmbER, SemDist, BERTScore were different, (C) metrics were contrasted with opposing predictions of the better hypothesis (e.g.  $WER_{(hypA)} > WER_{(hypB)}$  but  $CER_{(hypA)} < CER_{(hypB)}$ ).

Table 1 illustrates how hypotheses were matched with concrete examples.

**Table 1.** Detail of some stimuli choice criteria with examples. The  $\epsilon$  symbol represents a missing word.

Category	Metrics information	Reference	Hypothesis A	Hypothesis B
(A)	WER =	et on découvre les spectateurs and they discover the spectators	$\epsilon$ on découvre les spectateurs $\epsilon$ they discover the spectators	et on découvre les spectacles and they discover the show
(A)	CER >	sur la vie politique on the political life	$\epsilon$ la vie politique $\epsilon$ the political life	c' la vie politique t's the political life
(A)	SemDist >>	c' est à paris it's at paris	$\epsilon$ est à paris $\epsilon$ is at paris	c' est appau $\epsilon$ it's atpau $\epsilon$
(B)	WER = ; SemDist >	encore du rock still rock	corps du rock body of rock	encore du rok still rok
(C)	WER $\neq$ BERTscore	où les passions sont si vives where passions are so vivid	$\epsilon$ les patients sont si vive $\epsilon$ the patients are so vivid	où les patients sont si vifs where the patients are so lively

## 4 HATS data set

### 4.1 Corpus description

The HATS data set includes 1,000 references, each with two different erroneous hypotheses generated by different ASR systems. The preferred choice of 143 human evaluators for each 50 reference-hypotheses triplets is recorded in this data set, resulting in a total of 7,150 annotations. Each triplet is evaluated by at least 7 participants.

To assess the level of agreement between raters, we calculate Fleiss' Kappa, which yields a value of 0.46. In 82% of the triplet cases, the agreement (See Equation 1), is found to be at least 71.4%. Furthermore, in 60% of the triplet cases, the agreement reaches a minimum of 85.7%. This shows that the task is difficult, but that humans are still capable of determining a hypothesis as the best.

### 4.2 Methodology to evaluate metrics

Our method for evaluating metrics involved calculating the proportion of instances where both human annotators and the metric selected the same hypothesis as the best option.

Subjects were not allowed to determine that the two hypotheses were equal. However, it is certain that there are cases where one hypothesis cannot be chosen and the subject chooses randomly. Since the number of annotators for each triplet is 90% of the time odd, there will still be a winning hypothesis due to chance. One strategy to overcome this problem may be to take into account only the cases where there is a consensus. In this study, we calculate a human agreement that corresponds to a percentage indicating consensus. This is calculated according to the following formula:

$$\frac{\max(A, B)}{A + B} \quad (1)$$

where  $A$  is the number of humans who select one hypothesis, and  $B$  is the number of humans who select the other one. When agreement is weak, agreement is close to 50%, and if all humans agree on the same hypothesis, agreement is 100%. A filter can be applied on the data set according to three values of agreement: **100%** (keep only triplets where all subjects agree), **70%**, or **0%** (no filter applied); which corresponds to 371, 819 and 1000 utterances respectively. The 70% threshold was chosen in order to have consistent annotator agreement even if not all participants answer in the same way [28]. Taking the predictions of the metrics as a starting point, we calculate the number of times that humans chose the best hypothesis based on the evaluated metric.

## 5 Evaluation of ASR metrics from human perspective

Table 2 presents the results obtained by each metric according to the number of times they agree with human perception. Without surprise, the higher is the human agreement, the higher are the metrics performances. Unlike the results of previous studies [21], our study found that CER aligns more closely with human perception than WER. This divergence might be attributed to the use of written text as a reference in our perceptual experiment, rather than audio, or to intrinsic linguistic variations between French and English (French orthography contains a high number of silent letters compared to English).

It is interesting to note that at phoneme level, PER performs well, better than WER and CER despite the fact that humans have made their choices based on text alone. It shows that humans seem to consider how sentences sound even while reading. This is especially true if sentences are contrasted with a reference.

Although hypotheses selected based on BERTScore using BERT-base-multilingual perform 8% better than those chosen with SemDist Sentence multilingual, it would be premature to conclude that the BERTScore strategy is superior for evaluating the quality of transcripts as both metrics use different embeddings. When comparing these metrics with the same embeddings, SemDist outperforms BERTScore using CamemBERT-base embeddings while SemDist has a similar performance with BERTScore using CamemBERT-large. This suggests that some embeddings are more optimized for specific metrics.

On the 70% and 0% agreement level, WER have performances close to a random choice. This is due to the fact that in our data set, many cases present hypotheses with the same WER, and equal predictions are considered as a failure of the metric since humans are able to faithfully select one hypothesis. Furthermore, we can observe that



**Table 2.** Performance of each metric according to their human agreement. **Full** means that no filter on agreement were applied on data set. The number in parentheses indicates the percentage of times the metric gave the same score to both hypotheses.

<b>Agreement</b>	<b>100%</b>	<b>70%</b>	<b>0% (Full)</b>
Word Error Rate	63% (23%)	53% (28%)	49% (28%)
Character Error Rate	77% (17%)	64% (21%)	60% (22%)
Embedding Error Rate	73% (12%)	62% (16%)	57% (17%)
BERTScore BERT-base-multilingual	84% ( 0%)	75% ( 1%)	70% ( 1%)
BERTScore CamemBERT-base	81% ( 0%)	72% ( 0%)	68% ( 0%)
BERTScore CamemBERT-large	80% ( 0%)	68% ( 0%)	65% ( 0%)
SemDist CamemBERT-base	86% ( 0%)	74% ( 0%)	70% ( 0%)
SemDist CamemBERT-large	80% ( 0%)	71% ( 0%)	67% ( 0%)
SemDist Sentence CamemBERT-base	86% ( 0%)	75% ( 0%)	71% ( 0%)
SemDist Sentence CamemBERT-large	90% ( 0%)	78% ( 0%)	73% ( 0%)
SemDist Sentence multilingual	76% ( 0%)	66% ( 0%)	62% ( 0%)
SemDist FlauBERT-base	65% ( 0%)	62% ( 0%)	59% ( 0%)
Phoneme Error Rate	80% (14%)	69% (16%)	64% (17%)

SemDist using FlauBERT-base embeddings performs worse than CER. This highlights the importance of carefully selecting embeddings and evaluating them on data sets like HATS before drawing conclusions about system performances at a semantic level. Based on our human-oriented data set, the best metric is SemDist using Sentence CamemBERT-large, which can be explained by the fact that this metric is based on embeddings specifically trained to maximize the similarity between sentences with similar meanings. It is worth noting that a large amount of annotated data is necessary to use these embedding-based metrics.

## 6 Conclusion and Perspectives

In this study, automatic evaluation metrics applied to transcriptions coming from different ASR systems were compared to human evaluation of different erroneous hypotheses according to one written reference. Our results show that SemDist with Sentence-BERT evaluates transcripts in a way that seems acceptable for human raters. If Sentence-BERT is not a possible option, BERTScore seems to be the second best option. This metric is more stable than SemDist on BERT embeddings. Nevertheless, if possible, metrics should be evaluated through data sets comprising also human annotations such as HATS.

Although these new evaluation methods are interesting in the context of ASR, the advantage of WER and CER metrics lies in their computational low-cost and interpretability of the score. Therefore, the next step could be to develop metrics that correlate with human perception while remaining interpretable.

As future work, an additional study could be conducted by replicating the current experiment using an audio reference instead of a textual reference, so that subjects do not have character information. This approach would enable us to examine any variations and if CER is still considered as better than the WER in a multimodal setting.

## Limitations

The HATS data set is not necessarily representative of all kind of errors nor the most common because errors were selected applying strict criteria. In order to evaluate the representativeness of this data set, additional analyses with respect to the kind of errors that occur in each system’s transcriptions have to be carried out.

Furthermore, conclusions drawn from this data set may be specific to the French language and may not generalize to other languages. Adding and comparing similar data sets in other languages would help to better understand the performance of metrics and human evaluations across different languages.

## Ethics Statement

The aim of this paper is to propose a new method for evaluating speech-to-text systems that better aligns with human perception. However, the inherent subjectivity of transcription quality means that if we optimize systems to correlate only with the perception of the studied population, it could be inequitable if this perception does not generalize to the rest of the population.

## References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* **33**, 12449–12460 (2020)
2. Bañeras-Roux, T., Rouvier, M., Wottawa, J., Dufour, R.: Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. In: *Interspeech 2022* (2022)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Transactions of the association for computational linguistics* **5**, 135–146 (2017)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*. pp. 4171–4186 (2019)
5. Esteve, Y., Bazillon, T., Antoine, J.Y., Béchet, F., Farinas, J.: The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news. In: *International Conference on Language Resources and Evaluation (LREC)* (2010)
6. Evain, S., Nguyen, M.H., Le, H., Boito, M.Z., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., 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)
7. Favre, B., Cheung, K., Kazemian, S., Lee, A., Liu, Y., Munteanu, C., Nenkova, A., Ochei, D., Penn, G., Tratz, S., et al.: Automatic human utility evaluation of ASR systems: Does WER really predict performance? In: *INTERSPEECH*. pp. 3463–3467 (2013)
8. Freitag, M., Rei, R., Mathur, N., kiu Lo, C., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., Martins, A.F.: Results of WMT22 Metrics Shared Task: Stop Using BLEU—Neural Metrics Are Better and More Robust. In: *Proceedings of the Seventh Conference on Machine Translation, Abu Dhabi. Association for Computational Linguistics* (2022)

9. Freitag, M., Rei, R., Mathur, N., Lo, C.k., Stewart, C., Foster, G., Lavie, A., Bojar, O.: Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In: Proceedings of the Sixth Conference on Machine Translation. pp. 733–774 (2021)
10. Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.F., Mostefa, D., Choukri, K.: Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In: International Conference on Language Resources and Evaluation (LREC). pp. 139–142 (2006)
11. Galliano, S., Gravier, G., Chaubard, L.: The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In: Tenth Annual Conference of the International Speech Communication Association (2009)
12. Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., Quintard, L.: The repere corpus: a multimodal corpus for person recognition. In: International Conference on Language Resources and Evaluation (LREC). pp. 1102–1107 (2012)
13. Gordeeva, L., Ershov, V., Gulyaev, O., Kuralenok, I.: Meaning Error Rate: ASR domain-specific metric framework. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 458–466 (2021)
14. Grave, É., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
15. Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., Galibert, O.: 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). pp. 114–118 (2012)
16. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhota, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **29**, 3451–3460 (2021)
17. Itoh, N., Kurata, G., Tachibana, R., Nishimura, M.: A metric for evaluating speech recognizer output based on human-perception model. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
18. Juang, B.H., Rabiner, L.R.: Hidden Markov models for speech recognition. *Technometrics* **33**(3), 251–272 (1991)
19. Kafle, S., Huenerfauth, M.: Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility. pp. 165–174 (2017)
20. Kim, S., Arora, A., Le, D., Yeh, C.F., Fuegen, C., Kalinli, O., Seltzer, M.L.: Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. In: Proc. Interspeech 2021. pp. 1977–1981 (2021). <https://doi.org/10.21437/Interspeech.2021-1929>
21. Kim, S., Le, D., Zheng, W., Singh, T., Arora, A., Zhai, X., Fuegen, C., Kalinli, O., Seltzer, M.: Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric. In: Proc. Interspeech 2022. pp. 3978–3982 (2022). <https://doi.org/10.21437/Interspeech.2022-11144>
22. Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D.: FlauBERT: Unsupervised Language Model Pre-training for French. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 2479–2490 (2020)
23. Le, N.T., Servan, C., Lecouteux, B., Besacier, L.: Better Evaluation of ASR in Speech Translation Context Using Word Embeddings. In: Proc. Interspeech 2016. pp. 2538–2542 (2016). <https://doi.org/10.21437/Interspeech.2016-464>

24. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., De La Clergerie, É.V., Seddah, D., Sagot, B.: CamemBERT: a Tasty French Language Model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219 (2020)
25. Mathur, N., Wei, J., Freitag, M., Ma, Q., Bojar, O.: Results of the WMT20 metrics shared task. In: Proceedings of the Fifth Conference on Machine Translation. pp. 688–725 (2020)
26. Mdhaffar, S., Estève, Y., Hernandez, N., Laurent, A., Dufour, R., Quiniou, S.: Qualitative evaluation of asr adaptation in a lecture context: Application to the pastel corpus. In: INTER-SPEECH. pp. 569–573 (2019)
27. Nam, S., Fels, D.: Simulation of Subjective Closed Captioning Quality Assessment Using Prediction Models. *International Journal of Semantic Computing* **13**(01), 45–65 (2019)
28. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on Multimedia information retrieval. pp. 557–566 (2010)
29. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF, IEEE Signal Processing Society (2011)
30. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.C., Yeh, S.L., Fu, S.W., Liao, C.F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R.D., Bengio, Y.: SpeechBrain: A general-purpose speech toolkit (2021), [arXiv:2106.04624](https://arxiv.org/abs/2106.04624)
31. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
32. Vasilescu, I., Adda-Decker, M., Lamel, L.: Cross-lingual studies of ASR errors: paradigms for perceptual evaluations. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 3511–3518 (2012)
33. Wang, Y.Y., Acero, A., Chelba, C.: Is word error rate a good indicator for spoken language understanding accuracy. In: 2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721). pp. 577–582. IEEE (2003)
34. Zhang\*, T., Kishore\*, V., Wu\*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020), <https://openreview.net/forum?id=SkeHuCVFDr>