Building a Dataset for French Accent Classification Evaluation: Are We There Yet?

Diandra Fabre[1], Mathieu Avanzi[2], François Portet[1]
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2]U. Neuchâtel, Switzerland
{diandra.fabre, francois.portet}@univ-grenoble-alpes.fr, mathieu.avanzi@unine.ch

Abstract

Current evaluation practices in speech processing systems often overlook the diversity of spoken accents, leading to significant performance disparities across speaker groups. This issue largely comes from biases and imbalances in training corpora, and is further compounded by the scarcity of open-source datasets suitable for evaluating accent variability in French. To address this gap, we extend the CFPR dataset with explicit accent labels, providing a new benchmark for assessing the robustness of speech technology systems across diverse French accents. We additionally conduct a perceptual study with 87 human participants to evaluate the reliability and interpretability of these labels. Using this resource, we evaluated an eight-class French accent classifier trained on Common Voice data. The first results highlight both the complexity of automatic French accent recognition in low-resource settings, and the difficulty for French-speakers to perceive all the linguistic variabilities in French-speaking countries.

Keywords: accent recognition, benchmark, regional accent, perceptual studies, diversity, french

1. Introduction

The recent success in speech processing systems is largely due to the development of novel modeling techniques and the availability of extensive training data. Despite these advancements, a significant limitation persists in current evaluation practices: they often fail to adequately account for speech diversity . This diversity manifests across several dimensions, including gender (partially reflected in speaker fundamental frequency), speech register (tied to vocabulary choice), and accent (involving distinct acoustic features). Consequently, speech technology systems exhibit substantial performance disparities across speaker groups (Aman et al., 2013; Chang et al., 2024; Herron et al., 2025). For example, performance in Automatic Speech Recognition (ASR) drops for under-represented dialects, such as African American Vernacular English (AAVE), indicating that even Self-Supervised Learning (SSL) pre-training on large datasets does not fully mitigate these differences (Chang et al., 2024). These shortcomings are largely attributed to biases in the ASR training corpora. The scarcity of open-source datasets suitable for evaluating these diverse characteristics, particularly in non-English languages like French, exacerbates this problem, highlighting the need for new evaluation resources.

In this paper, we introduce the extension of the Corpus du Français Parlé de nos Régions (CFPR) corpus (Avanzi et al., 2019) with accent labels for the evaluation of speech technology systems on diverse French accents. The study also introduces a perception test involving 87 human participants designed to understand how these labels can be the most robustly acquired. This dataset was subsequently used to evaluate a French accent classifier that was trained on Common Voice.

French is spoken by over 320 million people across five continents, naturally resulting in a wide array of accents (e.g., regions of France, North Africa, Quebec). However, this accent diversity is underrepresented in current models since Self-Supervised Learning (SSL) models of French are typically (pre-)trained with less than 1% of non-standard French accent data (Parcollet et al., 2024).

In sociolinguistics, an accent can be defined as a set of pronunciation features that affect vowels, consonants, or prosody, and that make it possible to identify the profile of the person who uses them (Candea, 2021). An accent can be associated with a region, an age category, or a social group. In this paper, we will choose to focus on accent variability across regions.

A primary challenge in gathering accented data stems from the fact that existing datasets are often constructed with different target objectives, which leads to disparate categorization and labeling methods. For instance, Fair-Speech (Veliche et al., 2024) considers broad categories like L1 (native) versus L2 (non-native), and Sonos (Sekkat et al., 2024) uses home country instead of fine-grained accent labels. The labeling methods are also highly varied, with Common Voice and Fair-Speech relying on self-reported tags, while Sonos uses geographic proxies. Furthermore, most of these datasets are not conversational, hindering the speech variations that can be found in spontaneous speech.

To acquire accent labels, another method is to engage native speakers to perceptually rate speech

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5711–5721
11–16 May 2026.

in their own language. While perceptible variations still remain across generations in mainland France, significant variations are observed between mainland France, its overseas territories, and other French-speaking regions. A study by Woehrling and de Mareüil (2006) showed that people from mainland France generally distinguish only three broad accents: Southern, Northern, and Eastern French. It is yet to be investigated whether French-speaking people can perceptually distinguish these three broad mainland accents alongside the overseas varieties.

The contributions of this paper are the following :

- Extension of the CFPR dataset with regional accent labels.
- Perceptive test on a pool of 87 participants to evaluate the difficulty of this classification task for humans.
- Implementation and evaluation of a French accent classifier on eight classes trained on a dataset collected based on geographical criteria.

The remainder of this paper is organized as follows: Section 2 presents the state-of-the-art in accent classification and ressources. Section 3 details the annotation process and the perception study. Section 4 presents the French accent classifier, Section 5 the results, followed by a conclusion in Section 6.

2. Related Works

Speech corpora are numerous for English (Veaux et al., 2017; Demirsahin et al., 2020; Wang et al., 2024); however, across the literature, accent classification has relied on varying taxonomies (Veaux et al., 2017; Wang et al., 2024; Zhong et al., 2025)(Zuluaga-Gomez et al., 2023) or on unchecked labeling schemes such as the Common Voice self-declaration (Ardila et al., 2020). In contrast, accent-annotated speech corpora for non-English languages are much less frequent.

In 2007, Cappeau and Gadet (2007) observed, "There is no very large corpus of spoken French, and in particular, there has been no institutional will in France that would have led to the creation of a large oral corpus," while mentioning a high number of initiatives to collect data. Today, the situation has changed with the convergence of initiatives in the humanities that have collected large amounts of highly informative linguistic data, made available via open portals such as CoCoON or ORTOLANG, and with the emergence of deep learning, which encourages major players to collect and distribute speech data extracted from the web, such as the VoxPopuli corpus (Wang et al., 2021) or Mozilla's initial Common Voice. Although there are specialized corpora for specific varieties of French, such as the CEREALES Quebec French corpus (Maison et al., 2023), the Valibel Belgian French corpus (Francard et al., 2002), or Sub-Saharan African accented French (noa, 2003; Zanon Boito et al., 2022), there is a lack of a French accent corpus that provides both representative geographic and demographic coverage of French accents and the necessary sociolinguistic metadata for deep analytical study.

Regarding automatic accent classification, a recent review by Jassim and Abdulmohsin (2025) shows that most of the studies on accent classification focus on English variations. Some other studies focus on Arabic variations or South-East Asian variations. The research on accent classification is mainly driven by the need to improve Automatic Speech Recognition performances. For instance, Zuluaga-Gomez et al. (2023) introduced CommonAccent, an open-source benchmark for accent classification based on Mozilla Common Voice. Two large pretrained acoustic models were fine-tuned to classify accents in four languages: English, German, Spanish, and Italian. They achieved an accuracy from 68.5 to 99.0 according to the language and provided the first standardized baseline for accent classification across multiple languages. This demonstrates the effectiveness of self-supervised acoustic representations for accent classification. The more recent AccentBox (Zhong et al., 2025) builds on CommonAccent and achieved a 0.56 F1 score on unseen speakers over 13 accents of English.

As for speech embeddings, the most recent techniques rely on using representations from models such as WavLM, Wav2Vec2, or MMS-LID-256 on classification tasks. While WavLM is based on English speech, MMS-LID-256 is multilingual and 10 times bigger than WavLM.

To our knowledge, the first attempt at automatic French accent classification was proposed in a recent paper from Voxlect (Feng et al., 2025). In this study, French accents are divided into four classes: Switzerland/Belgium/Germany, Africa, Canada, France and French overseas territories being merged as a single accent class. Two different datasets were used: CommonVoice-fr (Ardila et al., 2020) and African Accented French (noa, 2003). Results of their classication model based on MMS-LID-256 reached 86.4 accuracy and 70.6 macro-F1 for French.

While Voxlect (Feng et al., 2025) represents a first effort in automatic French accent classification, its partitioning methodology is questionable. First, the authors did not specify if one given speaker could be present in different data splits (train/validation/test), which poses a risk of speaker

leakage and may lead to an overestimation of the model generalization capacity. Second, European, African, and Canadian varieties form identifiable but sometimes overlapping clusters within the classification results. While the authors acknowledge differences between French spoken in different countries, they omitted to take into account the significant differences inside France, both mainland and overseas.

3. Benchmarking accent recognition for French

3.1. Defining accent classes

Research into the perception of French accents indicates that native European French speakers were generally capable of distinguishing three main accent varieties of French (Woehrling and de Mareüil, 2006). A more recent work (Avanzi and de Mareüil, 2019) on eight European varieties showed the following different perceptual groupings: Alsace, Belgium, and Switzerland clustered together, North-West of France and Southern France (including Corsica).

Given our interest in the French accent both within and outside Europe, and building upon this perceptual evidence, we considered eight distinct classes. We acknowledge that these categories contain important inter-class variability, as they may embrace multiple regions or even multiple countries that present significant internal differences. Furthermore, some varieties of French are not covered (e.g., Italy). However, we relied on the perception research from Woehrling and de Mareüil (2006) for the initial classification of French European varieties and then applied this framework to the different French-speaking areas globally. The classes are defined in Table 1.

3.2. The CFPR corpus

The Corpus du Français Parlé de nos Régions (CFPR) (Avanzi et al., 2019) is an open-access linguistic resource developed at Sorbonne Université, designed for the study of regional and social variation in the French-speaking world. It comprises a collection of audio recordings. The corpus includes 186 speakers, each contributing one recording, along with detailed metadata. This metadata for each speaker includes: gender, year of birth, country and region of origin, current residence, recording date, recording location, and level of French. The metadata specifically provides both the speaker's birthplace and their residence at the time of the recording. The recordings were conducted as interviews with people living in various places around the world, from Côte d'Ivoire to Algeria, the main part originating from mainland France. This corpus is thus well-suited for the analysis of regional accents.

3.3. Gold annotation

The gold annotation[1] was obtained via concerted labeling. Two authors (one from Central France and one from Southern France) listened and annotated all recordings collaboratively, achieving consensus after multiple passes. They had a posteriori access to the full speaker metadata (childhood and current locations at time of recording) to resolve ambiguous cases and contacted a third person in a couple of cases.

When comparing the childhood location (origin) to the final accent label, 24 out of 186 of the recordings exhibited a mismatch. The specific divergences were: Central France (8 instances) finally labeled as Eastern and Northen France (5 instances) or unknown (3 instances); Canadian French (1 instance) labeled as unknown; Eastern and Northen France (6 instances) labeled as Centre; Southern France (7 instances) labeled as Centre (6 instances) and Eastern and Northen France (1 instance) and Pacific Area (1 instance) and North Africa (1 instance) both labeled as Centre. When comparing the current location to the final accent label, 60 out of 186 of the recordings exhibited a mismatch, showing that current location is not predictive of the persived accent.

In the remainder of this paper, we will refer to the labels resulting from this concerted labeling process as the "gold labels".

3.4. Human perception experiment

The online experiment concerning the perception of French accents was conducted between end September and mid-October 2025, and was implemented using the PsyToolkit platform (Stoet, 2010, 2017). The study was presented to participants having as a long-term objective to measure the fairness with which Automatic Speech Processing (ASP) systems treat the diversity of French accents. The experiment was completed online over approximately ten minutes, with no risk presented to participants, as their identity, microphone, and camera were not used. The participants were recruited in the immediate vicinity of the lab and the professional network of the experimenters. A very good command of French was required. Participants were instructed to use a physical keyboard in a quiet environment.

The records were preprocessed using diarization of each of the recordings to separate the speakers from the interviewer using pyannote (Bredin, 2023).

[1] https://zenodo.org/records/18848970

Table 1: French accent class definition

Class name	Geographical zones
Canadian French	Canada, Louisiana, Saint Pierre and Miquelon.
Caribbean	Guadeloupe, Martinique, French Guiana, Haiti…
Central France	All regions of France that are not part of Southern, Eastern, or Northern France.
Eastern & Northern France	Alsace, Luxembourg, Belgium, Switzerland, Pas-de-Calais, Moselle…
North Africa	Tunisia, Algeria, Morocco + Lebanon…
Pacific and Indian Area	Madagascar, Mauritius, Réunion Island, New Caledonia, Vanuatu…
Southern France	Basque Country, Occitania, Aquitaine, Provence, Corsica.
Sub-Saharan Africa	All French-speaking African countries south of the Sahara (*e.g.* Gabon, Congo, Ivory Coast, Niger, Cameroon…)

Table 2: Distribution of speakers' origin in the CFPR corpus

Origin region	Speakers	Percentage
Canadian French	6	3.23
Caribbean	7	3.76
Central France	75	40.32
Eastern and Northen France	31	16.67
North Africa	23	12.37
Pacific Area	5	2.69
Southern France	34	18.28
Sub-Saharan Africa	5	2.69
Total	**186**	**100.00**

For the evaluation task, we then selected the two longest utterances from each speaker, resulting in 372 segments for the participants to evaluate, with an average duration of 11.05 seconds (median of 11.4 seconds, min 5.1, max 13.1).

The experiment was structured into three distinct phases. First, a classical questionnaire was administered to collect demographic data. This was followed by a training phase, during which participants were trained on 5 controlled accent recordings to establish a baseline. Finally, the main annotation phase was executed, during which participants were asked to assign an accent to 25 distinct speech recordings into one of the eight accent classes. They were also encouraged to use the "I don't know" class each time they hesitated due to the task's acknowledged complexity. Each participant had 30 seconds to answer after hearing the speech record.

3.5. Results of the experiment

The experiment was performed by 87 participants. 47.8% of them were women, 44.8% men and 7.5% did not assign themselves as either of these genders. 22.7% were within 18-24-year-olds, 43.9% 25-34 yrs, 16.7% 35-44, 12.1% 45-54, 4.5% 55-64 and 1.5% more than 65 years old. Over all participants only four declared not to be native or C2 level. Regarding accent, most participants (46.3%) declared an accent of the center of France, 9% from the north/east, 14.9% from the south, 6% from North Africa and 1.5% from Caribbean. 22.4% did not select any accent, mostly to declare not having any accent.

On the training period, participants had an average guess rate of $0.5^{\text{SD}=0.2}$ (SD = standard deviation) and an average answer time of $13.5^{\text{SD}=4.2}$ seconds. The time-out rate was of 4%.

A total of 2,053 votes were collected at the end of the experiment. After excluding 308 votes categorized as "I don't know," 1,745 exploitable votes remained for analysis. Each recording received a median of 4 votes (1 min, 13 max). For each recording, the majority vote was computed to establish the initial labels. For the 56% of recordings that received more than three votes, the median Majority Ratio (the proportion of votes for the majority class) was 0.6 (0.2 min, 1 max). 19% of all recordings resulted in a tie for the majority class.

The decision for each speaker was determined solely by the best majority vote over the two recordings of each speaker. When comparing these majority labels against the "Gold labels", 55 out of 186 recordings mismatched. The overall agreement of the majority vote with the Gold labels demonstrated a low to moderate Krippendorff's Alpha (Krippendorff, 2019; Marzi et al., 2024) (0.62) and Cohen's Kappa (0.62).

Figure 1 exhibits the confusion matrix between the perceived accents and the gold labels. The recall for North Africa $(.82)$, North America (1), and Sub-Saharan Africa (1) shows that they where easily identified by the majority votes. Southern France (0.79) and the Caribbean (0.71) showed moderate identification accuracy, while Center $(.72)$ and North east $(.62)$ demonstrated great inter-confusion among raters. Pacific area $(.4)$ was poorly identified.

To assess whether aggregation metrics could serve as a reliable proxy for prediction quality, the dataset was partitioned into three subsets. High Confidence was defined by records having more than three total votes AND a majority ratio ≥ 0.75 of all votes. Moderate Confidence required more than three total votes AND a ratio of ≥ 0.5 AND the majority vote matching the speaker's origin. All other recordings were classified as Low Confidence.

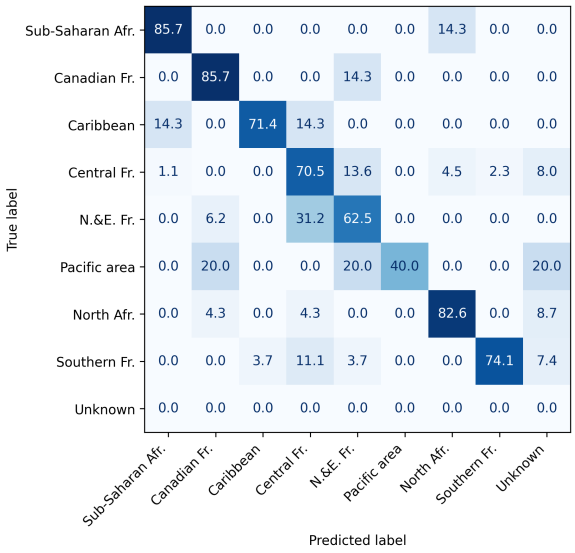


Figure 1: Perceptive test results on CFPR. Rows show true classes, columns predicted classes. Darker blue indicates a higher number of samples.

This process resulted in a distribution of 69 high, 32 moderate, and 85 low confidence recordings.

A subsequent analysis of inter-rater agreement across these subsets underscored the difficulty of the task. The full votes dataset showed extremely low agreement with a Krippendorff's Alpha of 0.266. Even after removing the low confidence votes (leaving 1,025 votes), the Alpha only rose to 0.345. When considering only the high confidence votes (651 votes remaining), the Alpha reached 0.453.

These findings lead to a clear conclusion: accent classification is a very challenging task, particularly for differentiating accents from closely related geographical areas (North East vs Centre) or for regions less represented in mainland (Pacific area), a difficulty exacerbated by the limited 10-second duration in average of the speech signals. While the speaker's childhood location (origin) serves as a potentially useful initial proxy for accent classification, the data demonstrates it cannot be fully relied upon since 13% of them were corrected by concerted labeling, indicating that perceived accent and geographical origin frequently diverge.

4. Automatic classification of French accents

The results of our perceptual study confirm that accent classification remains a challenging task for human listeners. This limitation motivates the exploration of automatic classification, to investigate whether an automatic approach can perform this task more reliably.

4.1. Methodology

Following experiments of Voxlect from Feng et al. (2025), we perform audio classification using a pre-trained WavLM model (`microsoft/wavlm-base`) and pre-trained MMS-lid 256 (`facebook/mms-lid-256`) as feature embeddings.

As the authors of Voxlect, we obtained significantly lower scores fine-tuning WavLM model than fine-tuning MMS-lid 256 model. We decided to focus on this last model. It is a fine-tuning of 1B parameters Massively Multilingual Speech (MMS, `facebook/mms-1b`) on a task of language identification (256 languages, including French, English, German). We freeze all model weights and replaced the classification layer with ours. The model is fine-tuned for single-label classification, with 8 classes as output. All audio files were resampled to 16 kHz.

To handle class imbalance, we use Focal Loss and weighted random sampling. Focal Loss gives more importance to difficult examples. We also use a custom balanced Trainer, which ensures that each batch contains a representative number of samples for each class, while oversampling underrepresented classes. We train with the AdamW optimizer and a cosine learning rate scheduler with warmup. The model is evaluated at the end of each epoch and the best model is saved after 10 epochs.

Accuracy and F1-scores are the metric used for evaluating the performance of the models.

4.2. Datasets

As there is no balanced accent dataset for French, we collected datasets from different sources, with variable accent annotation quality.

4.2.1. CommonVoice 22.0

Common Voice[2] defines itself as a free, open source platform for community-led data creation. Since 2021, it is possible for users to self-determine their accent to label their speech production. However, there could be as many labels as there are persons.

To construct our dataset, we only selected audio data with clear accent identification. We gathered more than 200 different labels, the most occurring one being "Français de France" (French of France). However, this label covers multiple accents and can not be interpreted easily. As described by Candea (2021), accents are inherently unstable and subjective, as pronunciation features may be perceived differently by listeners or change in prominence over time. There is a notion of a "neutral" pronunciation with French, considered accent-free and

[2]https://commonvoice.mozilla.org/fr

serving as the social standard, while any deviation is labeled an accent, often with negative social connotations. This hierarchy places the accent around Paris and center of mainland France area as the norm and the neutral way of speaking. Based on these observations, we made the hypothesis that all occurrences labeled as "French of France" belonged to this accent class.

We chose explicit labels showing clear accent identification for each one of our eight classes. We excluded all labels where French was not a spoken language of the country such as "French of Germany". From all the available data, we kept 37555 recordings with a length between 3 and 13 seconds, for a total of 63.05 hours, and got content for the eight classes as shown in Table 4. This data were split between train and validation subsets, using the speaker identification to make sure none of the speakers of the validation subset were in the training subset.

Total durations vary importantly between classes. We observe three dominant classes in term of size: Canadian French, Central France, and Eastern & Northern France.

4.2.2. Dataset augmentation

In an effort to balance all classes, we gathered data from other open source datasets. We added 18305 recording segments between 3 and 13 seconds for a total of 28.35 additional hours. Data are presented in Table 3 and organized as follow:

- PFC [3] (Durand et al., 2002): Southern France (Aix-Marseille, Douzens, Lacaune), North Africa (Chlef in Algeria), Pacific Area (Réunion Island)
- African accented French (noa, 2003): Sub-Saharan Africa (Cameroon, Chad, Congo, Gabon and Niger)
- Assemblée Nationale : speeches of French National Assembly deputees, three from the Caribbean area (DR, JPN, CB, EC, JW, JVC), four from the Pacific area (MS, ET, AB, FG), one from Southern France (JC).

4.3. Results from automatic classification

We trained two models, one using the CommonVoice dataset only (M_cv22), and a second one adding the augmentation dataset (M_aug). For both models, only the training subset changed, while the validation set remained the same. We tested these two models on the CFPR dataset presented in Section 3.2, on a total of 6599 audio segments extracted. For each experiment, we aggregated all accent predictions from a given speaker and kept the most frequently predicted label at the speaker level. This reduces intra-speaker variability and ensures that evaluation reflects consistent accent perception rather than isolated utterance-level fluctuations.

Results from automatic classification are displayed in Figures 2 and 3, for the models respectively trained on the CommonVoice dataset and on the augmented dataset. Per-class accuracy and F1-score are also provided in Table 5.

Using only CommonVoice as training dataset, we observe low performance for classes representing the Caribbean and the Pacific Area. These classes are poorly represented in our dataset. However, while being one of the largest classes in our dataset, Eastern & Northern France also displays underperforming results, right above chance. Figure 2 shows that Central France, one of the largest classes, is also showing as a wrong prediction for half of the classes.

When training on the augmented dataset, we can see that even minor data augmentation on the smallest classes can significantly improve the results, especially for the Caribbean (1.23 h to 2.94 h) and the Pacific Area (2.02 h to 5.19 h). The impact on Southern France given the data increase (2.15 h to 6.90 h) is also positive (about 13% accuracy increase with a stable F1-score). Finally, North Africa class results are consistent before and after augmentation (while increasing from 0.80 h to 4.11 h).

Table 3: Augmentation repartition from open-source datasets

Dataset	Label	Duration
African Accented French	Sub-Saharan Africa	15.42 h
PFC	Pacific area	2.52 h
	North Africa	3.31 h
	Southern France	4.29 h
Assemblée Nationale	Caribbean	1.71 h
	Pacific area	0.65 h
	Southern France	0.46 h
Total		**28.35 h**

Table 4: Dataset durations per class, with Common Voice 22.0 (CV22.0) and with augmentation (in hours), for cumulated train + val dataset

Label	CV22.0	Augmented
Sub-Saharan Africa	2.29 h	17.71 h
Canadian French	15.19 h	15.19 h
Caribbean	1.23 h	2.94 h
Central France	17.29 h	17.29 h
Eastern & Northern France	22.07 h	22.07 h
Pacific Area	2.02 h	5.19 h
North Africa	0.80 h	4.11 h
Southern France	2.15 h	6.90 h
Total	**63.05 h**	**91.41 h**

[3] http://www.projet-pfc.net

We observe a slight increase in accuracy and a light drop in F1-score. Performance for Central France and Eastern & Northern France drops significantly after augmentation. These classes show confusion with multiple other regions, including the Pacific Area, North Africa, and Southern France. These results testify once again that these two classes might not capture inter-class variability and cover both area and accent differences with characteristics similar to those of other classes. In Figure 3, the Pacific Area class shows a strong pull in predictions from dominant classes, indicating potential over-generalization.

Table 5: Per-class accuracy and F1-score for mms-lid-256 model fine-tuned on CommonVoice 22.0 and on augmented dataset.

Class	M_cv22	M_aug	M_cv22	M_aug
	Accuracy		F1-score	
Sub-Saharan Africa	**68.20**	51.61	0.626	**0.661**
Canadian French	**81.00**	76.00	0.271	**0.616**
Caribbean	6.83	**53.41**	0.116	**0.463**
Central France	**54.32**	11.75	**0.565**	0.195
Eastern & Northern Fr.	**15.54**	2.86	**0.220**	0.053
Pacific Area	5.15	**52.06**	0.034	**0.063**
North Africa	21.88	**25.48**	**0.331**	0.302
Southern France	45.33	**58.27**	0.411	**0.432**
Global accuracy	**40.98**	**25.04**		
Macro-avg recall	**37.28**	**41.43**		

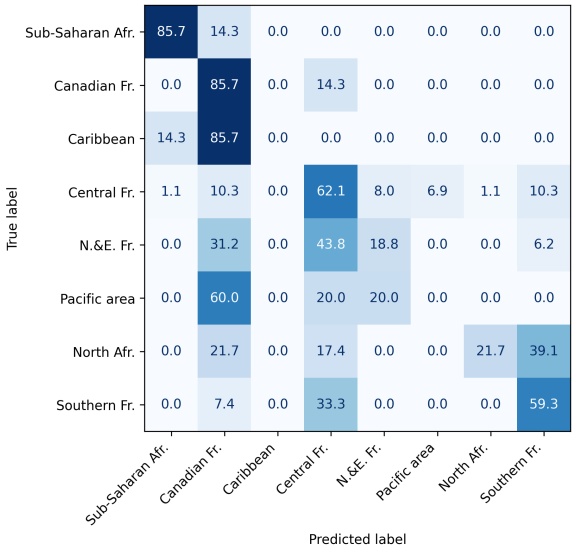


Figure 2: Confusion matrix for the M_cv22 model trained on CommonVoice 22.0 and tested on tested on CFPR. Rows show true classes, columns predicted classes. Darker blue indicates a higher number of samples.

To compare with current state-of-the-art approach, Figure 4 reports the results obtained by the French model of Feng et al. (2025). We compared Voxlect with the performance of our models by joining different class results. Africa includes

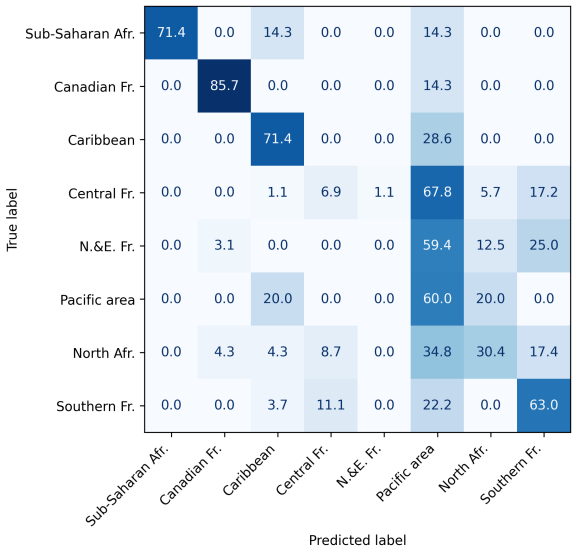
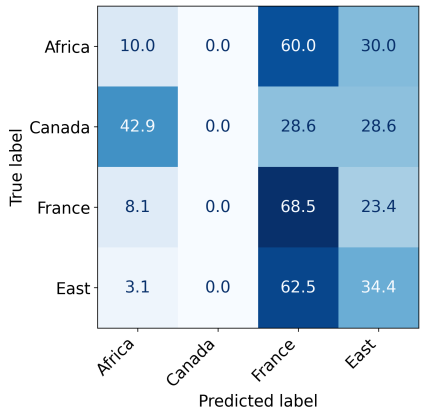


Figure 3: Confusion matrix for M_aug model trained on Commonvoice 22.0 and additional data, and tested on tested on CFPR. Rows show true classes, columns predicted classes. Darker blue indicates a higher number of samples.

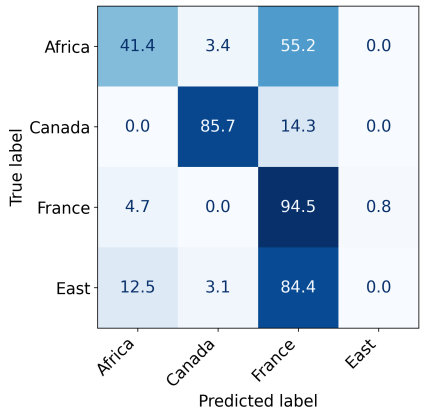
Sub-Saharan Africa and North Africa, France regroups Caribbean, Central France, Pacific Area and Southern France. This Figure thus compares an 4-class architecture with our 8-class architecture. Both models are tested on CFPR dataset, where for each speaker, we only kept the most frequent label for all utterances. The Voxlect study reports that the main confusion for the lowest-performing class occurs with the 'France' class. Tested on CFPR dataset, we observe that Africa subset and Canada subset have the lowest accuracy. In our results, the aggregated 'Africa' class performs better than in Voxlect. For our model M_aug, the aggregation of finer-grained regional labels of 'France' into one category leads to significantly better performance compared to Voxlect. Classification accuracy for the 'East of France' class is the lowest, while Voxlect manages to reach a 34.4% accuracy.

Despite good results on the 'France' label for the coarse classes, this hides the discriminative power of the models on the fine-grained regional accents in France. This expresses the need for a richer variability and accent repartition over French-speaking regions.

Based on these experiments, we can hypothesize that classes such as Central France and Northern & Eastern France are too dominant. These classes might include various noisy data or even incorrectly labeled accents from the related CommonVoice categories. Northern & Eastern France gathers accents from different regions and even different countries such as Belgium and Luxembourg. That might include too much variability.



(a) Results for Voxlect model on CFPR dataset with 4 labels.



(b) Results for M_aug model on CFPR dataset with 4 labels

Figure 4: Confusion matrices for both Voxlect and our model tested on CFPR dataset with 4 labels: Africa, Canada, France, East.

Finally, we can hypothesize that the Pacific Area could also be a too diverse class and should be divided into more specific regions. Indeed, while languages in the geographical area from Madagascar to Tahiti come from the same Austronesian families (Keenan and Chung, 2017), their native speakers exhibit pronunciation differences when speaking French. We based our class split on the availability of data and the perceived proximity of pronunciations. However, the datasets are not distributed evenly across the geographic areas they represent. For instance, North Africa class augmentation comes from recordings realised in the same city, sometimes in noisy conditions.

Next steps should include a study of literature on phoneme-level acoustic differences inside the classes to refine their distribution. We used strategies to balance classes in the training steps of the models, but should focus in the next step on increasing the smallest classes to increase the number of speakers. Refining the class definitions based on linguistic criteria could make the model distinctions more meaningful and improve overall performance.

5. Discussion and Conclusion

The Voxlect study (Feng et al., 2025), which utilized the MMS-LID-256 architecture, reported a macro F1-score of 0.707 when classifying French accents into four broad classes (Switzerland/Belgium/Germany, Africa, Canada, and France and its overseas territories). This result was achieved on short utterances. In contrast, our approach aimed for a more granular analysis by classifying accents into eight distinct classes (reflecting a finer sociolinguistic partitioning). Using the same MMS-LID architecture, our system achieved a lower classification accuracy of 40.98% across these eight classes, which is close to some studies in English (0.56 F1 score on 13 accents of English (Zhong et al., 2025)).

This performance difference highlights several current limitations in research on automatic accent classification. Our low accuracy when doubling the number of classes underscores the difficulty in separating more specific regional varieties of French. This also illustrates the lack of resources for training such models. Acquiring such a dataset has proven to be difficult. Research on French native listener perception in Europe (Avanzi and de Mareüil, 2017) found that, even when tasked to choose among France, Switzerland, and Belgium, human accuracy averaged only 60%. By contrast, our experiment based on a majority vote led to an accuracy of 70.4%. Despite a very low inter-annotator agreement, this is far better than the machine learning approach.

When access to childhood origin is available, using such information as a proxy for accent led to a higher accuracy of 87.1%. This high score is likely influenced by the CFPR's focus on recruiting people from specific regions. However, when the current location is used (a common proxy based on ZIP code to guess a remote speaker's demographics), the accuracy falls to 68%, illustrating the unreliability of this proxy for accent attribution.

This reveals that humans struggle in this task and might explain why automatic accent classification systems still fail on this task. This also shows that access to metadata, even if not perfect, is of high importance to achieve more reliable labelling.

Our future research will focus on advancing French automatic accent classification. We aim to collect and curate more high-quality datasets to address current resource scarcity, and extend the proposed evaluation dataset. Furthermore, we plan to systematically evaluate a broader range of speech technologies, such as Automatic Speech Recognition and speaker identification models. Ultimately, we plan to apply accent classification to

enable large-scale sociolinguistic studies.

Acknowledgments

We would like to warmly thank all the participants of the study for their precious feedbacks and their help in diffusing the perception test, and also Andréa Unn-Toc for her precious expertise on Caribbean area accents. This research has been partially funded by the French National Research Agency under the France 2030 program, reference ANR-23-IACL-0006 (AugmentIA and "AI & Language" chairs), ANR-23-IAS1-0001 (The Pantagruel project) and ANR-22-CE23-0013 (E-SSL project). This work was performed using HPC resources from GENCI at IDRIS under the allocations 2024-A0171014633, and 2025-A0191013801 on the Jean Zay supercomputers.

6. Bibliographical References

Frédéric Aman, Michel Vacher, Solange Rossato, and François Portet. 2013. Analyzing the performance of automatic speech recognition for ageing voice: Does it correlate with dependency level? In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pages 9–15.

Mathieu Avanzi and Philippe Boula de Mareüil. 2017. Identification of regional french accents in (northern) france, belgium, and switzerland. *Journal of Linguistic Geography*, 5(1):17–40.

Mathieu Avanzi and Philippe Boula de Mareüil. 2019. Peut-on identifier perceptivement huit accents régionaux en français européen? la réponse des sciences participatives. *Glottopol: Revue de sociolinguistique en ligne*, 31:1–21.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Maria Candea. 2020. Accents et styles de prononciation au prisme de la norme du français. In Alexandra Cunita and Coman Lupu, editors, *Norma şi uz în limbile romanice actuale*, volume 31 of *Romanica*, pages 53–65. Editura universitatii din Bucuresti.

Maria Candea. 2021. Accent. *Langage et société*, (HS1):19–22.

Paul Cappeau and Françoise Gadet. 2007. Document 3: Où en sont les corpus sur les français parlés? *Revue française de linguistique appliquée*, (1):129–133.

Kalvin Chang, Yi-Hui Chou, Jiatong Shi, Hsuan-Ming Chen, Nicole Holliday, Odette Scharenborg, and David R Mortensen. 2024. Self-supervised speech representations still struggle with african american vernacular english. In *Proc. Interspeech 2024*, pages 4643–4647.

Jacques Durand, Bernard Laks, and Chantal Lyche. 2002. La phonologie du français contemporain (pfc): usages, variétés et structure. *Romanistische Korpuslinguistik romance corpus linguistics*, pages 93–106.

Tiantian Feng, Kevin Huang, Anfeng Xu, Xuan Shi, Thanathai Lertpetchpun, Jihwan Lee, Yoonjeong Lee, Dani Byrd, and Shrikanth Narayanan. 2025. Voxlect: A speech foundation model benchmark for modeling dialects and regional languages around the globe. *arXiv preprint arXiv:2508.01691*.

Felix Herron, Solange Rossato, Alexandre Allauzen, Benoit Favre, and François Portet. 2025. Speaker Group Encoding in Self-supervised Speech Recognition Models. In *Text, Speech, and Dialogue (TSD)*, volume 16029 of *Lecture Notes in Computer Science*, pages 121–132, ERLANGEN, France. Springer Nature Switzerland.

Maliha Jahan, Yinglun Sun, Priyam Mazumdar, Zsuzsanna Fagyal, Thomas Thebaud, Jesus Villalba, Mark Hasegawa-Johnson, Najim Dehak, and Laureano Moro Velazquez. 2025. Faist: A benchmark dataset for fairness in speech technology. In *Proc. Interspeech 2025*, pages 1343–1347.

Sarah Jassim and Husam Ali Abdulmohsin. 2025. Accent classification using machine learning techniques: A review. *International Journal of Computer Information Systems and Industrial Management Applications*, 17:421–451.

Edward L Keenan and Sandra Chung. 2017. The austronesian languages.(asia-pacific linguistics.).

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: a user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, et al. 2024. Lebenchmark 2.0: A standardized, replicable

and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, 86:101622.

Kathy Reid and Elizabeth T. Williams. 2023. Common Voice and accent choice: data contributors self-describe their spoken accents in diverse ways. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–10, Boston MA USA. ACM.

Chloé Sekkat, Fanny Leroy, Salima Mdhaffar, Blake Perry Smith, Yannick Estève, Joseph Dureau, and Alice Coucke. 2024. Sonos voice control bias assessment dataset: A methodology for demographic bias assessment in voice assistants. In *LREC/COLING*.

Gijsbert Stoet. 2010. Psytoolkit: A software package for programming psychological experiments using linux. *Behavior research methods*, 42(4):1096–1104.

Gijsbert Stoet. 2017. Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1):24–31.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-speech dataset. *arXiv preprint arXiv:2408.12734*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Cécile Woehrling and Philippe Boula de Mareüil. 2006. Identification of regional accents in french: perception and categorization. In *INTERSPEECH*.

Jinzuomu Zhong, Korin Richmond, Zhiba Su, and Siqi Sun. 2025. Accentbox: Towards high-fidelity zero-shot accent generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. 2023. CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. In *INTERSPEECH 2023*, pages 5291–5295. ISCA.

7. Language Resource References

2003. African Accented French. Type: dataset.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Mathieu Avanzi, André Thibault, and François Delafontaine. 2019. Corpus du français parlé de nos régions (cfpr).

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. 2020. Open-source multi-speaker corpora of the english accents in the british isles. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6532–6541.

Michel Francard, Geneviève Geron, and Régine Wilmet. 2002. La banque de données valibel: des ressources textuelles orales pour l'étude du. *Romance corpus linguistics: corpora and spoken language*, 1:71.

Lucas Maison, Thomas Soulas, and Marie-Jean Meurs. 2023. Cereales: a new dataset of quebec french accented speech with applications to speech recognition. In *Proc. Interspeech*, pages 4058–4062.

Christophe Veaux, Junichi Yamagishi, Korin MacDonald, and et al. 2017. CSTR VCTK Corpus: English Multispeaker Corpus for CSTR Voice Cloning Toolkit.

Wenbin Wang, Yang Song, and Sanjay Jha. 2024. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech. In *Proc. Interspeech 2024*, pages 1365–1369.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael

Rouvier, and Yannick Estève. 2022. Speech resources in the Tamasheq language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2066–2071, Marseille, France. European Language Resources Association.