Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Vers l'apprentissage de modèles auto-supervisés de reconnaissance automatique de la parole plus équitables sans a priori démographique

Laura Alonzo-Canul[1] Benjamin Lecouteux[1] François Portet[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
laura.alonzo-canul@univ-grenoble-alpes.fr,
benjamin.lecouteux@univ-grenoble-alpes.fr, françois.portet@imag.fr

RÉSUMÉ

Malgré des avancées importantes dans le domaine de la Reconnaissance Automatique de la Parole (RAP), les performances de reconnaissance restent inégales selon les groupes de locuteurs, ce qui pose des problèmes d'équité. Bien qu'il existe des méthodes pour réduire ces inégalités, elles dépendent de ressources externes au signal vocal, telles que des modèles de locuteur (speaker embeddings) ou des étiquettes démographiques textuelles, qui peuvent être indisponibles ou peu fiables. Dans ce travail, nous proposons une méthode pour améliorer l'équité dans la RAP qui ne dépend d'aucune de ces ressources. Notre approche utilise une méthode de clustering non supervisé à partir de représentations acoustiques classiques, auto-supervisées et hybrides. Nos expériences avec CommonVoice 16.1 démontrent que les modèles entraînés sur les clusters découverts améliorent les performances des groupes démographiques désavantagés tout en conservant des performances compétitives et en utilisant deux fois moins de données d'entraînement.

ABSTRACT

Towards training fair self-supervised automatic speech recognition models without demographic labels

Research in Automatic Speech Recognition has highlighted important performance disparities across demographic groups of speakers, raising fairness concerns. Although efforts have been made to reduce the disparities, current approaches rely on resources other than the speech signal, such as speaker embedding models or textual demographic labels, which might not be readily available or scarce in self-supervised learning contexts. In this work, we propose a method to improve ASR fairness that does not rely on either of these resources. Our approach uses unsupervised clustering to leverage classical, embedded and hybrid classical-embedded speech representations, which can be computed without speaker-specific models and need no extra metadata. Experiments on English CommonVoice 16.1 show that models trained on the discovered clusters improve the performance of demographic groups in disadvantage while maintaining strong performance. Furthermore, clusters learned from classical representations outperform clusters of self-supervised representations and achieve near-baseline performance with half the training data.

MOTS-CLÉS : équité, apprentissage auto-supervisé, reconnaissance automatique de la parole.

KEYWORDS: fairness, self-supervised learning, automatic speech recognition, low-level speech features.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 780–790.

1 Introduction

Much like it did for text-based and image-based applications powered by artificial intelligence, recent years have seen speech-based applications become ubiquitous : from the comfort of our own homes, we live in continuous interaction with smart devices that grant us immediate access to speech-based services. Like in text and image processing too, speech processing has benefited from the rise of self-supervised learning (SSL), a machine learning paradigm that seeks to extract general and robust representations of data without the need for labels. Speech systems that rely on this paradigm have systematically seen performance gains across different tasks and domains largely due to an ever-growing variety of network sizes, network architectures, and the open release of very large data resources. However, although representations learned by these networks have indeed been shown to be meaningful, they have also been shown to be harmful (Martin & Tang, 2020; Mengesha *et al.*, 2021; Liu *et al.*, 2022), exhibiting a wide range of socioeconomic and demographic biases learned from data when applied to end-user applications.

Training and releasing language models without care or accountability for the biases they contain does not only enable the perpetuation of such biases but also directly hampers the development of fair and inclusive technologies. For instance, research in Automatic Speech Recognition (ASR) has shown that users from demographics known to be disadvantaged by a system might feel the need to adjust their speech to resemble that of a demographic they believe is best recognized in order to obtain better performance (Mengesha *et al.*, 2021). Scenarios such as this one, which occur as a direct consequence of deploying heavily biased systems into production environments, can compromise the feedback received by these systems ultimately leading to the development of technologies that are even more biased and less robust to out-of-distribution inputs.

Yet, in the context of ASR, only a handful of works have targeted the study of system biases (Koenecke *et al.*, 2020; Meyer *et al.*, 2020; Maison & Estève, 2023; Kulkarni *et al.*, 2024; Lin *et al.*, 2024) and even fewer attempts have been made to try to alleviate them (Boito *et al.*, 2022; Maison & Estève, 2023; Lin *et al.*, 2024). Moreover, these works rely heavily on the usage of textual demographic labels to either balance out training corpora across demographic groups or over-represent demographic groups in disadvantage, both of which have been shown to be insufficient to achieve fairness (Garnerin *et al.*, 2021; Maison & Estève, 2023).

In our view, several sensible and perhaps obvious reasons exist as for why *textual demographic labels are an unsuitable resource to train fair self-supervised speech processing models*. Demographic label categories are often too coarse to be used on their own (gender categories alone disregard the differences between speakers of different age or different accent), too granular to be applied together (data groups built out of couples of gender and age categories are often of insufficient size for training), or simply too complex to be defined (accent categories are often let to the user to describe, resulting in no consensus across different samples and even different datasets). Furthermore, demographic label categories do not necessarily correlate to voice features and often overlap each other (the voices of young males are more similar to that of young females than to adult males). Aside from the inherent challenges of obtaining manual annotations for these labels—such as their high cost, time-consuming nature and privacy-related issues—it is hard to imagine that these labels will ever be descriptive enough to account for speech variations and subtleties among speakers, which is essential to build representative demographic groups. Finally, manually annotated demographic labels also introduce unnecessary biases into the developing pipelines (Hutiri & Ding, 2022), which further hinders the creation of fairer models.

In this work, we present an approach for improving the demographic fairness of a self-supervised ASR model that does not rely on external speaker embedding models or textual demographic labels. Instead, our approach leverages classical, self-supervised and hybrid classical-embedded representations, which can be computed from the speech signal directly or extracted from the same self-supervised model. Our proposed approach also leverages unsupervised clustering of these features, which further supports our claim that training fair ASR models can be attained without the use of demographic labels.

The contributions of this work towards reaching fairness in ASR are the following :

1. We assess the usability of different speech representations, either extracted from a self-supervised model or directly from the speech signal, to build training subsets to improve fairness in ASR performance across demographics groups.

2. We explore the application of unsupervised clustering as the algorithm to build these training subsets and we empirically show that classical features provide a representation that is both more effective than self-supervised representations for the purpose of finding emerging training groups to improve ASR fairness.

2 Related work

Recent research in ASR has highlighted significant performance disparities between demographic groups. For instance, with respect to their ethnicity, authors in (Mengesha *et al.*, 2021) found that native African-American English speakers experience word error rates up to twice as high as White standard American English speakers. With respect to the spoken dialect, authors in (Martin & Tang, 2020) found ASR systems are two to four times less accurate at correctly inferring habitual "be"—a key morpho-syntactic feature of African American English—compared to non-habitual "be" in standard English. In terms of both dialect and gender, authors in (Tatman, 2017) found Youtube's automatic captifor ASR system performed worse overall for female speakers and speakers of highly accented variants, such as Scottish English, when evaluating the system on five different regional English dialects. Apart from these findings, the presence of socio-economic and socio-demographic biases in speech processing can be tracked down to much more than just end-system performance. As shown by (Hutiri & Ding, 2022) in their study of the VoxCeleb Speaker Recognition challenge, biases exist and can stem from every stage of a speech processing pipeline. Works in the field seeking to improve ASR performance across demographic groups have approached the task by either trying to build relevant data groups for training using demographic labels as guidance (Boito *et al.*, 2022; Maison & Estève, 2023) or trying to learn demographic group differences directly within the model (Dheram *et al.*, 2022; Veliche & Fung, 2023). Our work is similar to those in the first category in the sense that we aim to develop an effective data-driven approach to reduce the disparities, and related to those in the second category by the application unsupervised clustering in our respective pipelines. Our work, however, differs from previous efforts in that we do not rely on demographic labels to build training groups but we leverage classical, as well as self-supervised representations, as features to find emerging data groups in an unsupervised fashion. To the best of our knowledge, we have found no evidence of any other work in ASR that explores and compares the effectiveness of these representations for building training clusters with the aim of reducing demographic disparities.

3 Method

Our pipeline for training fairer ASR models builds on three consecutive steps : computing representations from speech data, applying unsupervised clustering over these representations and fine-tuning a self-supervised model for ASR on cluster-derived training subsets. We further describe the details of each of these steps below.

3.1 Representation extraction

The first step of our pipeline focuses on representation extraction. Specifically, we extract various types of representations from the training data to be used for training group discovery via unsupervised clustering. As introduced in Section 1, our goal is to investigate whether classical features or strong contextualized embeddings from a self-supervised model can serve as effective representations for group discovery—and ultimately contribute to reducing ASR performance disparities across demographic groups. To this end, we experiment with five different types of representations : two based on classical acoustic features, and three derived from self-supervised models.

Classical features commonly used in speech processing include Mel-Frequency Cepstral Coefficients (MFCCs) and the Log Power Spectrum (LPS) of the signal. In this work, we construct two custom speech representations based on these features : one MFCC-based and one LPS-based.

The MFCC-based representation is obtained by extracting 13 MFCCs from the raw speech signal, along with their first- and second-order derivatives (delta and delta-delta features). These 39-dimensional vectors are then averaged along the time axis, resulting in a single 39-dimensional representational vector per utterance.

The LPS-based representation is derived by applying the Short-Time Fourier Transform (STFT) with an FFT length of 2048. From the resulting complex-valued spectrogram, we retain only the positive frequency components (from 0 to the Nyquist frequency). We compute the squared magnitude to obtain the power spectrum, which is then converted to a logarithmic scale in decibels. As with the MFCC representation, the log power values are averaged over time, yielding a final 1025-dimensional vector representing the average log power per frequency bin.

For the self-supervised representations, we extract WAV2VEC 2.0 (Baevski *et al.*, 2020) contextualized embeddings, which are commonly obtained by pooling the output of the model's final hidden layer, resulting in a 1024-dimensional feature vector.

Classical acoustic features, such as MFCCs and LPS, are grounded in well-established signal processing principles—such as perceptual scaling or spectral energy distribution—that offer interpretable and compact characterizations of speech. In contrast, features learned through self-supervised models are highly expressive but lack direct interpretability. To investigate a *hybrid* representation combining classical signal-derived features and those learned via large-scale self-supervised models (such as WAV2VEC 2.0 , which has approximately 317 million parameters and is pre-trained on 960 hours of speech), we also experiment with an additional type of representation : self-supervised embeddings fine-tuned to predict classical features. This approach—using acoustic feature regressors to guide the training of deep speech encoders—was first introduced in (Pascual *et al.*, 2019). Following this idea, we modify the WAV2VEC 2.0 architecture by adding a lightweight feedforward regressor head (256 hidden units with PReLU activation) on top of the encoder, and train the model to predict MFCC or

LPS feature vectors using a mean squared error loss. We extract contextualized embeddings from these fine-tuned models in the same way as from the original WAV2VEC 2.0 model to obtain the two hybrid 1024-dimensional vector representations.

Beyond serving as a bridge between classical and self-supervised paradigms, this approach also allows us to investigate whether incorporating information from classical representations into a self-supervised model—by explicitly learning to reconstruct these features—can enrich the contextualized embeddings and help reduce ASR performance disparities across demographic groups. Our hypothesis is that aligning the learned representations with interpretable, human-engineered features may enhance the model's sensitivity to relevant phonetic and spectral cues, helping it better capture variation in speech—particularly from underrepresented demographic groups—and, therefore, improving ASR robustness across diverse speaker populations.

In summary, in this step we extract five distinct sets of representations from the training data : (1) MFCC-based features (raw-m13d), (2) Log Power Spectrum-based features (raw-lps), (3) WAV2VEC 2.0 embeddings (w2v2-embed), (4) WAV2VEC 2.0 embeddings fine-tuned to predict MFCC features (w2v2-ft-m13d), and (5) WAV2VEC 2.0 embeddings fine-tuned to predict LPS features (w2v2-ft-lps).

3.2 Finding emerging training groups through unsupervised clustering

The second step of the pipeline consists of clustering each set of representations to discover emerging training groups in an unsupervised fashion.

We use K-means as our clustering algorithm and run it over our five sets of representations k times with $k \in [2, 8]$. We choose this range for k deliberately, to enforce meaningful data constraints and ensure future interpretability. Specifically, we choose k to be substantially smaller than the total number of unique speakers in the dataset ($36,775$), while being congruent with the number of available demographic categories : at least equal to the number of gender labels (2) and no greater than the number of age labels with sufficient annotations (8 out of 9). Selecting a small number of k promotes the creation of clusters that represent broader, more diverse speaker groups and reduces the risk of creating many small, potentially noisy clusters of data that could impair downstream task stability and diminish statistical power for analyzing ASR performance disparities.

3.3 Training ASR models on the discovered groups

The last step of the pipeline consist of training self-supervised models for ASR based on the clusters discovered in the previous stage. We build cluster-specific subsets of the training data, where each subset corresponds to samples grouped according to the K-means -derived clusters. We then train a WAV2VEC 2.0 model for ASR on each of these subsets, using the standard Connectionist Temporal Classification (CTC) loss. We evaluate model performance on the original development set and report test set Word Error Rate (WER) scores in the upper part of Table 1.

To further validate our approach and, more precisely, to assess the impact of unsupervised clustering on training group discovery, we also trained a set of baseline models. The first baseline model (w2v2-baseline) is a WAV2VEC 2.0 model trained on the full original training set, without any cluster-based data selection. Additionally, we trained eight *random baselines*, where each model was trained on a randomly selected subset of data. The size of each subset corresponds to the size of the original

training set divided by a factor of k, with $k \in [2, 8]$. WER scores obtained by these random baseline models are reported in the lower part of Table 1.

3.4 Metrics of interest

In addition to overall WER, we examine performance gaps across demographic categories. Following a similar approach to (Dheram *et al.*, 2022), we compute the WER-Gap for each demographic category as :

$$\text{WER-Gap} = \frac{\text{Avg-WER}_{\text{bottom}} - \text{Avg-WER}_{\text{top}}}{\text{Avg-WER}_{\text{top}}}$$

Here, for each demographic category, bottom refers to the set of demographic labels with WER scores lower than the average, while top includes those with WER scores higher than the average.

Note that this score is always zero or negative. A more negative WER-gap value indicates a larger disparity, meaning that lower-WER demographic groups perform significantly better than higher-WER groups. A WER-Gap of zero reflects demographic fairness, meaning all groups have equal WER.

4 Experiment Details

4.1 Data

We use the English CommonVoice 16.1 dataset (Ardila *et al.*, 2020) for our experiments, whose training set contains over one million speech utterances with less than ten seconds of read speech, as well as manual demographic label annotations for three demographic categories : gender (three sub-categories), age (nine sub-categories) and accent. Due to a lack of consensus among speakers for the accent category (accent information was self-reported through an open-text field, resulting in inconsistent and non-standardized labels), only the gender and age categories are used for the quantitative evaluation of our approach.

4.2 Models

We use the wav2vec2-large checkpoint from the HuggingFace Hub for embedding extraction, fine-tuning embeddings on classical features and training for ASR. We implement clustering with K-means using scikit-learn. All our models and experiments are implemented with the HugginFace library. Our code is publicly available at `https://github.com/alonzocl/towards-fair-asr`.

model name	k	k_{id}	dev	test	female	male	other	no-g-label	teens	twenties	thirties	fourties	fifties	sixties	seventies	eighties	no-a-label	train %
Trained on cluster-based sets																		
w2v2-baseline	-	-	**16.7**	**19.7**	**19.5**	**21.3**	**28.5**	**19.5**	**22.4**	**22.8**	**19.8**	**16.6**	**16.9**	**14.5**	**14.5**	**13.3**	**19.5**	**100.0**
w2v2-embed	2	1	18.2	21.1	20.8	22.5	**23.6**	20.8	23.3	23.8	21.4	18.2	19.4	14.5	14.5	13.3	20.8	52.4
	3	1	18.5	21.6	20.4	23.3	25.7	21.4	23.4	24.5	21.9	19.4	18.8	**12.6**	15.5	13.3	21.4	43.9
	4	1	19.0	22.0	20.9	23.7	26.4	21.8	24.8	24.8	22.4	19.1	18.8	15.3	13.9	**6.7**	21.8	28.7
raw-lps	2	0	**17.4**	**20.3**	**19.7**	**21.8**	25.3	**20.2**	**22.2**	**23.1**	**20.7**	**17.4**	**17.8**	14.2	13.0	20.0	**20.2**	72.7
	3	2	18.0	21.0	20.2	22.7	24.3	20.8	23.2	23.7	21.9	18.8	18.9	14.8	13.3	16.7	20.8	56.0
	4	2	18.6	21.6	20.2	23.3	26.4	21.3	24.1	24.4	21.9	18.8	18.8	15.6	13.6	10.0	21.3	39.3
raw-m13d	2	0	17.7	20.8	20.3	22.3	27.4	20.6	23.1	23.2	21.5	18.8	18.7	14.8	**12.1**	20.0	20.6	62.0
	3	0	18.1	21.4	20.9	22.9	25.7	21.2	23.4	24.0	21.9	19.4	19.1	14.8	15.2	13.3	21.2	49.2
	4	0	18.6	21.7	21.2	23.4	27.1	21.4	23.9	24.8	22.1	18.8	19.8	15.0	13.0	**6.7**	21.4	41.1
w2v2-ft-lps	2	1	18.1	21.2	20.6	23.2	24.3	20.9	23.9	24.2	21.9	19.2	19.1	14.5	13.3	16.7	20.9	54.1
	3	1	18.3	21.5	20.9	23.3	25.7	21.3	24.3	24.5	21.6	19.2	20.2	15.0	13.9	10.0	21.3	43.4
	2	0	18.5	21.7	21.3	23.3	27.8	21.5	23.8	24.8	22.3	19.6	18.0	15.3	14.2	20.0	21.5	45.9
w2v2-ft-m13d	2	1	18.0	21.0	20.5	22.8	28.1	20.7	23.3	23.9	21.9	19.1	18.7	14.8	14.5	16.7	20.7	54.0
	2	0	18.4	21.7	21.0	23.4	27.1	21.4	24.1	24.6	22.3	19.2	19.0	15.0	15.8	13.3	21.4	46.0
	3	1	18.5	21.6	21.2	23.2	27.8	21.3	23.6	24.5	22.4	18.5	18.8	15.3	14.2	**6.7**	21.4	43.4
Trained on randomly sampled sets																		
	2	-	20.6	24.0	**22.98**	**25.52**	29.51	23.78	26.02	**27.01**	23.95	21.68	**20.13**	15.85	17.27	**16.67**	23.77	50.0
	3	-	**20.4**	**23.7**	23.54	25.53	**26.74**	**23.44**	**25.99**	27.12	**23.86**	21.26	21.72	**15.03**	15.76	**16.67**	**23.45**	33.3
	4	-	21.4	24.7	23.63	26.94	27.78	24.38	26.39	28.57	25.47	**21.16**	23.13	18.03	16.06	30.0	24.38	25.0

TABLE 1 – Test set WER scores of WAV2VEC 2.0 models fine-tuned for ASR using different sets of training data. The sets were constructed from K-means clusters (top) or random sampling (bottom), as explained in Section 3.3.

model name	k	k_{id}	# spk	% spk	female	male	other	no-g-label	teens	twenties	thirties	fourties	fifties	sixties	seventies	eighties	no-a-label
Trained on cluster-based sets																	
w2v2-baseline	-	-	36775	100.0	19.8	49.1	2.3	28.8	6.9	27.5	14.0	10.2	6.3	5.6	0.6	0.1	28.8
w2v2-emb	2	1	34190	93.0	19.3	48.5	2.6	29.5	8.3	30.1	13.3	8.7	5.3	4.3	0.5	0.1	29.5
	3	1	32444	88.2	20.0	49.1	2.3	28.6	6.9	28.0	14.2	10.0	6.2	5.5	0.5	0.1	28.6
	4	1	29509	80.2	19.6	48.4	2.5	29.5	7.8	29.4	13.6	9.0	5.4	4.6	0.5	0.1	29.5
raw-lps	2	0	35446	96.4	25.0	41.9	2.5	30.6	7.2	27.4	12.6	9.5	6.1	5.8	0.6	0.1	30.6
	3	2	33629	91.4	30.5	34.7	2.8	32.0	7.8	27.4	12.1	7.9	5.7	6.3	0.7	0.1	32.0
	4	2	30102	81.9	16.5	53.0	1.9	28.5	5.6	27.1	15.4	9.4	6.8	6.3	0.5	0.1	28.5
raw-m13d	2	0	28117	76.5	20.6	50.1	2.4	26.8	6.9	29.0	13.1	11.3	5.7	6.4	0.5	0.1	26.9
	3	0	28951	78.7	18.3	50.9	2.2	28.5	6.7	27.4	15.9	10.1	6.1	4.6	0.6	0.0	28.6
	4	0	26289	71.5	18.7	51.6	2.3	27.4	6.7	28.2	15.1	11.0	5.8	5.0	0.6	0.0	27.5
w2v2-ft-lps	2	1	34945	95.0	19.1	48.8	2.5	29.7	7.7	28.3	13.8	8.9	5.9	5.0	0.5	0.1	29.7
	3	1	32666	88.8	20.1	49.0	2.2	28.6	6.7	27.5	14.3	9.9	6.4	5.8	0.6	0.1	28.7
	2	0	30965	84.2	20.8	49.4	2.0	27.8	5.9	26.5	14.2	11.9	6.8	6.2	0.6	0.1	27.9
w2v2-ft-m13d	2	1	34944	95.0	19.1	48.8	2.5	29.7	7.7	28.4	13.8	8.9	5.9	5.0	0.5	0.1	29.7
	2	0	30965	84.2	20.8	49.4	2.0	27.8	5.9	26.5	14.2	11.8	6.8	6.2	0.6	0.1	27.9
	3	1	32683	88.9	20.1	49.0	2.2	28.7	6.7	27.5	14.3	9.9	6.4	5.8	0.6	0.1	28.7

TABLE 2 – Distribution of utterances across demographic labels for the training subsets used for the models in Table 1. The number of unique speakers per set (# spk), along with its proportion relative to the number of unique speakers in the original training set of CommonVoice 16.1 (% spk), is also reported.

5 Results and Discussion

5.1 ASR performances across demographic groups

Table 1 shows the recognition results per demographic category achieved on the CommonVoice 16.1 dataset using our proposed pipeline. Model names in the table correspond to the specific type of representation used consistently across all steps of the pipeline—representation extraction, clustering, and ASR fine-tuning—for the construction of each model. k and k_{id} are the number of clusters learned by the K-means algorithm during the clustering step, and the cluster id that was used to build the training set that obtained the reported ASR performance, respectively. 'train %' shows the portion of the training dataset assigned by the K-means to the respective cluster id. 'no-g-label' and 'no-a-label' denote utterances in CommonVoice 16.1 lacking gender and age annotations, respectively. Bold values highlight the lowest WER scores for each demographic group while blue values denote

performance improvements over the baseline. For simplicity, we only report WER scores for the top three models per type of representation in the top part of the table, plus results on our baseline.

According to the WER scores per demographic category obtained by our baseline model (w2v2-baseline), utterances spoken by female speakers are easier to transcribe than those spoken by males, but none of them are as hard as those spoken by speakers who identified themselves with the "other" gender label, the category that scored the highest WER among all demographic categories (28.5 WER). With respect to their age, utterances from speakers in their eighties achieved the best recognition scores (13.3 WER) while those from speakers in their twenties achieved the worst (22.8 WER).

When comparing overall performance across best models, models trained on clusters learned from embedded representations (w2v2-embed, w2v2-ft-lps, w2v2-ft-m13d) delivered almost identical test set performance (21.1, 21.2 and 21.0 WER, respectively), with WAV2VEC 2.0 embeddings fine-tuned on MFCCs taking a very slight lead over standard WAV2VEC 2.0 embeddings but with the later achieving a significant performance improvement on the "other" category, the hardest to transcribe from the corpora. We note, however, that all three models exhibited less than a two-point difference in WER relative to the baseline, despite having been trained on roughly half the amount of data. Our models trained on clusters learned from representations extracted directly from the speech signal (raw-lps, raw-m13d), rather than self-supervised embeddings, both achieved better performance than their embedded counterparts (w2v2-ft-lps, w2v2-ft-m13d). More interestingly, all of our models trained on clusters of non-standard WAV2VEC 2.0 representations (first row for all raw-lps, raw-m13d, w2v2-ft-lps and w2v2-ft-m13d) outperformed all top three models trained on clusters of standard WAV2VEC 2.0 embeddings (w2v2-embed), demonstrating that classical features offer a strong alternative to self-supervised representations for the purpose of improving fairness in ASR. The best performance among all of our models was achieved by a model trained on clusters learned from LPS features (raw-lps), which not only achieved lower WER scores consistently across most demographic groups in CommonVoice 16.1 compared to the other models, but also showed improved performance in four demographic categories, including the challenging 'other' category, compared to our WAV2VEC 2.0 baseline, a model trained on over one million speech utterances.

Table 3 shows the WER-Gap scores for the gender and age demographic categories. As seen on the Table, training on clusters reduces the gap between gender groups, regardless of the type of representation used during the clustering step. Additionally, the top model from each raw-lps, raw-m13d and w2v2-ft-m13d, also managed to reduce the gap between age groups with respect to the baseline. We plan to look more carefully into better ways to measure these trade-offs in future work.

5.2 Effects of data size, number of speakers and clusters

When examining the results in the lower section of Table 1, we observe that, as expected, all of our models outperformed their corresponding random baseline relative to the amount of training data used. In addition, unlike the models trained on subsets of data filtered by the unsupervised clustering step, none of the random baselines improved WER for any demographic category. With respect to the number of clusters learned by the K-means , we also found that exploring up to $k = 8$ was sufficient for our task and setup, as none of our models trained with one of $k \in [5, 8]$ clusters obtained better overall recognition performance.

Lastly, one of the most important effects we observed from our experiments is that, among all representations explored in this work, the one based on the LPS was the only one that allowed the

model	k	k_{id}	Gender Gap	Age Gap
w2v2-baseline	-	-	**-0.284**	**-0.300**
w2v2-emb	2	1	**-0.098**	-0.312
	3	1	-0.167	-0.361
	4	1	-0.166	-0.456
raw-lps	2	0	-0.180	-0.274
	3	2	**-0.140**	-0.281
	4	2	-0.176	-0.395
raw-m13d	2	0	-0.223	**-0.267**
	3	0	-0.148	-0.331
	4	0	-0.177	-0.471
w2v2-ft-lps	2	1	**-0.133**	-0.315
	3	1	**-0.140**	-0.410
	2	0	-0.198	**-0.262**
w2v2-ft-m13d	2	1	-0.230	**-0.272**
	2	0	-0.181	-0.300
	3	1	-0.201	-0.440

TABLE 3 – WER-gap per demographic category. The three models with the smallest gaps, plus the baseline, are highlighted in bold.

clustering algorithm to assign a maximum of both utterances and unique speakers to a single cluster, an effect that did not occur when clustering self-supervised representations.

6 Conclusion

In this work, we introduce a three step pipeline-approach for mitigating demographic group performance disparities in ASR by leveraging different types of representations, including ones defined by classical acoustic features and self-supervised models. Our findings demonstrate that classical features such as MFCCs and LPS provide a meaningful representation that can be exploited through unsupervised clustering to form emerging training groups that improve ASR fairness. Compared to self-supervised embeddings, these representations resulted in more balanced WER performance across demographic groups while maintaining competitive performance. The results of our research suggest that fairness in ASR can be enhanced through feature-driven clustering, reducing the reliance on potentially biased or insufficient demographic annotations. Our study contributes to the broader effort of developing more equitable speech technologies and encourages further exploration of feature-informed learning paradigms in self-supervised speech processing.

Acknowledgments

This work was supported by the ANR E-SSL project (N°ANR-22-CE23-0013) and used HPC resources from GENCI–IDRIS (A0171014633).

Références

ARDILA R., BRANSON M., DAVIS K., HENRETTY M., KOHLER M., MEYER J., MORAIS R., SAUNDERS L., TYERS F. M. & WEBER G. (2020). Common voice : A massively-multilingual

speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, p. 4211–4215.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

BOITO M. Z., BESACIER L., TOMASHENKO N. & ESTÈVE Y. (2022). A study of gender impact in self-supervised models for speech-to-text systems. *arXiv preprint arXiv :2204.01397*.

DHERAM P., RAMAKRISHNAN M., RAJU A., CHEN I.-F., KING B., POWELL K., SABOOWALA M., SHETTY K. & STOLCKE A. (2022). Toward fairness in speech recognition : Discovery and mitigation of performance disparities. *arXiv preprint arXiv :2207.11345*.

GARNERIN M., ROSSATO S. & BESACIER L. (2021). Investigating the impact of gender representation in ASR training data : a case study on librispeech. In M. COSTA-JUSSA, H. GONEN, C. HARDMEIER & K. WEBSTER, Éds., *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, p. 86–92, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.gebnlp-1.10.

HUTIRI W. T. & DING A. Y. (2022). Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, p. 230–247.

KOENECKE A., NAM A., LAKE E., NUDELL J., QUARTEY M., MENGESHA Z., TOUPS C., RICKFORD J. R., JURAFSKY D. & GOEL S. (2020). Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, **117**(14), 7684–7689.

KULKARNI A., KULKARNI A., TRANCOSO I. & COUCEIRO M. (2024). Unveiling biases while embracing sustainability : Assessing the dual challenges of automatic speech recognition systems. In *Interspeech 2024*.

LIN Y.-C., LIN T.-Q., LIN H.-C., LIU A. T. & LEE H.-Y. (2024). On the social bias of speech self-supervised models. *arXiv preprint arXiv :2406.04997*.

LIU C., PICHENY M., SARI L., CHITKARA P., XIAO A., ZHANG X., CHOU M., ALVARADO A., HAZIRBAS C. & SARAF Y. (2022). Towards measuring fairness in speech recognition : Casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6162–6166 : IEEE.

MAISON L. & ESTÈVE Y. (2023). Some voices are too common : Building fair speech recognition systems using the common voice dataset. *arXiv preprint arXiv :2306.03773*.

MARTIN J. L. & TANG K. (2020). Understanding racial disparities in automatic speech recognition : The case of habitual" be". In *Interspeech*, p. 626–630.

MENGESHA Z., HELDRETH C., LAHAV M., SUBLEWSKI J. & TUENNERMAN E. (2021). "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, **4**, 169.

MEYER J., RAUCHENSTEIN L., EISENBERG J. D. & HOWELL N. (2020). Artie bias corpus : An open dataset for detecting demographic bias in speech applications. In *Proceedings of the twelfth language resources and evaluation conference*, p. 6462–6468.

PASCUAL S., RAVANELLI M., SERRÀ J., BONAFONTE A. & BENGIO Y. (2019). Learning problem-agnostic speech representations from multiple self-supervised tasks.

TATMAN R. (2017). Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, p. 53–59.

VELICHE I.-E. & FUNG P. (2023). Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5 : IEEE.