# Discovering Bias in Dutch Automatic Speech Recognition by Clustering Interpretable Acoustic and Prosodic Features

**Kayleigh Jones**[1]
**Supervisors: Odette Scharenborg**[1]**, Jorge Martinez Castaneda**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Kayleigh Jones
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, Jorge Martinez Castaneda, Merve Gürel

# Abstract

Dutch State-of-the-art Automatic Speech Recognition (ASR) systems do not perform equally well for different speaker groups. Existing metrics to quantify this bias rely on demographic metadata, which is often unavailable. Recent advances in the field use machine learning to find groups of similar speakers instead. However, its black-box nature obscures the interpretability of resulting groups. This paper proposes an interpretable approach to bias discovery by clustering speakers based on acoustic and prosodic features. Different feature subsets were compared in their ability to find performance disparities in five ASR systems for two separate speaking styles. Results show that these feature sets can uncover bias approaching known disparities between demographic groups. While the effectiveness per feature set differed between the speaking styles, the most successful ones found significant disparities between clusters with diverse demographic compositions.

**Index Terms**: speech recognition, interpretability, bias, fairness, feature extraction

## 1. Introduction

Automatic Speech Recognition (ASR) is the process of transforming spoken words into text. An ASR system is typically based on a deep neural network and trained on a large variety of transcribed conversations. While becoming increasingly integrated into our daily lives, recent evidence shows that State-of-the-art (SotA) ASR systems do not perform equally well for different speaker groups [1, 2, 3, 4, 5]. For instance, children and nonnative speakers are less accurately recognized than native adult speakers in Dutch [4, 1, 5, 3]. Similarly, performance disparities have been found between dialect regions [1, 3]. This bias is typically quantified by comparing the Word Error Rates (WERs) between speaker groups. Research in fairness for ASR compares ASR performances for demographic speaker groups under the assumption that the speakers within a group share similar voice characteristics [4]. However, most databases offer limited demographic information, also called metadata, which complicates conventional bias quantification between demographic groups. This indicates a need for alternative ways to split speakers into groups that share global patterns of pronunciation. Not only would more databases qualify for bias estimation, but it additionally has the potential to identify new types of speaker groups that exhibit significant and meaningful biases.

Dheram et al. [6] proposed automatic cohort discovery by clustering speaker embeddings, enabling bias quantification while eliminating the need for demographic labels. However, these embeddings were extracted using a deep neural network [7], resulting in uninterpretable speaker groups ("cohorts", in Dheram et al.'s work). While the advancement enables automatic bias mitigation leading to higher accuracy [6], emerging views on fairness believe research on ASR performance should address inclusion as well [3]. The proposed approach by Dheram et al. does not address this aspect. Research into interpretable alternatives for cohort discovery is nonexistent to the best of my knowledge.

While fairness in ASR is a relatively nascent topic in academia [1, 6], the assumption that demographic groups exhibit different pronunciation patterns is grounded in decades of phonetic research. Ladefoged and Broadbent [8] attributed variation in speech to three different sources: (1) variation between phonemes; (2) physical differences; and (3) sociolinguistic variation. ASR systems aim to detect phonemic differences, i.e., distinguish which of all possible sounds in the language is being pronounced by the speaker, while ignoring physical and social characteristics [9]. However, the presence of known performance disparities [2, 4, 1, 5, 3] suggests that the remaining variability has an undesirable impact on the ASR system's acoustic model [4]. Bias due to age is linked to physical variation in the vocal tract length [10], while speakers with regional or nonnative accents exhibit sociolinguistic variation [9]. Therefore, I hypothesize that leveraging known age- and accent-related speech variability can aid in the detection of poorly recognized speaker groups for ASR models.

The aim of this paper is to investigate the effectiveness of language-specific acoustic and prosodic feature sets in the discovery of performance disparities between groups of similar-sounding speakers. The overarching goal is to eliminate the need for demographic information in bias quantification while preserving the interpretability of the key vocal characteristics of poorly recognized groups. Following a brief literature study on speech variability, 17 features were selected and extracted from a database of diverse speech. With known disparities between demographic groups from [5] as a baseline, different feature subsets were evaluated in terms of bias detection. The feature sets that discovered the highest performance disparities per speaking style were interpreted in more detail. I hope to advance the field of fairness in ASR by providing a first report on interpretable bias discovery and recommending possible improvements to the approach for future researchers.

## 2. Methodology

The approach consists of the following parts. First, features were extracted from a corpus of diverse Dutch speech containing two different speaking styles. Then, multiple combinations of the features were explored for a broader overview of how acoustic and prosodic features perform at finding bias in ASR. For every feature subset and speaking style, the feature sets were clustered into speaker groups, and ASR performance was calculated based on these.

### 2.1. Dutch Corpora

Feature extraction was done on the full recordings and transcriptions from the NL region in the JASMIN corpus [11]. The JASMIN corpus contains diverse speech from (1) Dutch children (DC); (2) Dutch teenagers (DT); (3) nonnative teenagers (NnT); (4) nonnative adults (NnA); and (5) Dutch seniors (DOA). It is divided into two different speaking styles: read (Rd) speech and Human-Machine Interaction (HMI), both of which were considered separately in this paper. Table 1 shows the number of speakers and hours of speech data per demographic group for each speaking style.

### 2.2. Feature Selection

Seventeen features known to vary between speakers of the Dutch language were derived from a diverse set of studies in phonetics and ASR fairness [10, 12, 13, 4, 14, 15, 16, 17]. The features extracted from the JASMIN data included 16 acoustic and one prosodic feature. Acoustic features are measurements on individual phonemes. Prosody is suprasegmental, thus covering variability in acoustic features across utterances [18].

Table 1: *Number of speakers and hours of speech data used per demographic group of the NL region in the JASMIN corpus. For DOA, #Spks is broken down by speaking style (Rd, HMI) due to differences in exclusion in the feature extraction step.*

| Group | #Spks | #Hrs (Rd) | #Hrs (HMI) |
|---|---|---|---|
| DC | 71 | 8:36:34 | 6:34:34 |
| DT | 63 | 6:30:05 | 4:29:15 |
| NnT | 53 | 7:51:19 | 4:52:29 |
| NnA | 45 | 7:43:04 | 7:29:43 |
| DOA | 68, 67 | 8:35:26 | 10:17:44 |

### *2.2.1. Mean Pitch of Phoneme Segments*

The mean pitch of a voice is usually higher for children than for teenagers [10, 12]. In older adults, pitch has been found to decrease [13]. The mean pitch was measured over all phoneme segments, to ensure measurements do not include utterances from the machine in Human-Machine Interaction (HMI).

### *2.2.2. Articulation Rate*

Speaking rate is often slower for older adults, as well as for children [4, 12] and nonnative speakers [14, 15, 16]. There are two ways to measure speaking rate. The speech rate can defined as the number of spoken phonemes divided by the total duration of the recording. The articulation rate can be calculated by dividing the number of spoken phonemes by the total duration of the recording excluding all pauses [15]. In the present research, the articulation rate was used due to the challenge of using speech rate on Human-Machine Interaction (HMI), where speakers pause for longer times to listen to the machine's responses.

### *2.2.3. Mean Vowel Formants and Durations*

Vowels can be represented quantitatively by their formant frequencies. Speech can essentially be defined as a source-filter combination [19]. The vocal folds are the source, producing a sound called the glottal tone. The mouth is the filter, which can be controlled by the speaker using, among other things, the tongue. The mouth acts as a resonating chamber where different frequencies are amplified depending on the position of the tongue. These frequencies are referred to as formants. The first two formants, F1 and F2, respectively correlate closely to tongue height and backness [9]. Thus, by measuring F1 and F2 of some vowel for each speaker, comparisons can be made between speaker groups' overall tongue positions for that vowel, known as its vowel quality.

Only vowels with significant variability between speakers or low ASR performance were selected. The focus of this research lies on interpretability of characteristics of speaker groups, thus a limited yet linguistically-motivated selection of features is assumed beneficial for the clarity and relevance of the results.

Due to their known socio-linguistic variation [17], mean vowel durations were measured in addition to the formant frequencies. The following five vowels were considered: (1) /ɛ/, expected to vary in F1 between different regions in the Netherlands [17]; (2) /ɑ/, expected to differ in F2 between regions in the Netherlands; (3) /ə/, a phoneme commonly misrecognized by ASR systems [4] that only occurs in unstressed syllables and was therefore not analysed in the phonetic research by Adank et al.'s work [20, 17]; (4) /ɔ/, found to be hard to recognize for ASR systems when pronounced by speakers from the North of the Netherlands [4] and has shown a longer average duration for Northern Standard Dutch (spoken in the Netherlands) and Southern Standard Dutch (spoken in Flanders) [17]; and (5) /u/, expected to differ in F2 between communities within the Netherlands [17].

The phonemes /ʏ/, /y/ and /œy/ are commonly misrecognized by ASR systems when pronounced by nonnative speakers [4]. Unfortunately, for each of these phonemes, at least 50 out of 300 speakers in HMI data never used it, thus insufficient measurements were be obtained for these speakers during feature extraction. To avoid excessive data exclusion, /ʏ/, /y/ and /œy/ were removed from the feature space instead. However, this suggests that HMI and read speech differ in word choice, which can provide future directions for research into feature relevance.

## 2.3. Feature Extraction

Feature extraction was done using Praat version 6.4.12 [21]. Due to the non-linearity of human hearing [22], formant features were measured in the perceptually relevant bark-scale[1] to avoid overemphasis of higher frequencies. Remaining configurations were left on Praat's default settings, with a formant ceiling[2] of 5500 Hz and a pitch range[3] of 50-800 Hz. The Praat manual recommends using these values for analysing female voices, but adjusting them for male speakers or children. Although the JASMIN corpus contains each of these speaker categories, demographic metadata is assumed absent throughout the present research. Therefore, frequency features may be less accurate for some speakers than for others. The resulting feature vectors per speaker were exported into CSV format with 10 decimals precision.

Speakers that did not pronounce at least one of the phonemes of the selected feature space were excluded from the research. This could happen when (1) the phoneme never occurred in the transcriptions; or (2) when the phoneme did occur, but Praat could not find enough formants. This lead to the exclusion of one speaker from the DOA group for HMI speech. For read speech, no data was excluded.

## 2.4. Data Pre-processing

No vowel normalization was applied to account for anatomical variation between speakers, despite formant features being selected exclusively for socio-linguistic variation. Lobanov's z-score transformation [23] can be applied to eliminate unwanted variability for Dutch vowels [9]. Unfortunately, it requires measurements of all vowels, which are not available for a large portion of the JASMIN data. Consequently, the features used in this paper may exhibit collinearity, which may impact the results due to increased importance of certain features when using Euclidean distances.

New temporal features were engineered by multiplying each speaker's vowel durations with their corresponding articulation rate. Unprocessed temporal features exhibited significant linear pairwise correlation, causing unwanted

[1] https://www.fon.hum.uva.nl/praat/manual/_hertzToBark_.html

[2] https://www.fon.hum.uva.nl/praat/manual/Sound__To_Formant__burg_____.html

[3] https://www.fon.hum.uva.nl/praat/manual/Intro_4_2__Configuring_the_pitch_contour.html

Table 2: *Features included in each of the five feature subsets used in the experiments. The possible features were mean pitch ("Pitch"), mean articulation rate ("Artic. Rate"), and mean duration ("Dur"), mean first and mean second formant frequencies ("F1" and "F2", respectively) of five Dutch vowels. Vowels are written in IPA notation. "Mean" was omitted from column names due to space limitations. Checkmarks ("✓") indicate that a feature was included in the feature set, and an empty space that it was not.*

| Feature Subset | Pitch | Artic. Rate | /ɛ/ | | | /ɑ/ | | | /u/ | | | /ɔ/ | | | /ə/ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **Dur** | **F1** | **F2** | **Dur** | **F1** | **F2** | **Dur** | **F1** | **F2** | **Dur** | **F1** | **F2** | **Dur** | **F1** | **F2** |
| **Pitch** | ✓ | | | | | | | | | | | | | | | | |
| **Artic** | | ✓ | | | | | | | | | | | | | | | |
| **Adank** | | | | ✓ | ✓ | | | ✓ | | | ✓ | | ✓ | ✓ | | | |
| **Feng+** | ✓ | ✓ | | | | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **All** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

increased feature importance. The way articulation rate was measured allowed for the transformation of phoneme durations into durations relative to the speaker's overall phoneme rate. These engineered features replace the durational features for the remainder of the research.

To ensure equal importance of each feature, min-max scaling from the *scikit-learn* library [24] was applied to the features before clustering. Before selecting a scaling method, I conducted the Shapiro-Wilk [25] test for normality from the *SciPy* library [26] on feature vectors from both speaking styles. Results indicated that the majority of the 17 extracted features is not normally distributed ($p < 0.05$). Consequently, min-max scaling was used for its simplicity in handling feature spaces with varying distributions. However, future research could explore alternative scaling or standardization methods when sufficient expert knowledge is obtained about the significance of the underlying variability of chosen features.

No dimensionality reduction was applied to the feature space. While techniques such as Principal Component Analysis (PCA) may prove beneficial for minimizing linear correlations in the feature space, it engineers new features at the cost of interpretability. Therefore, it does not align with the main focus of the research.

### 2.5. Feature Subsets

Five feature subsets were compiled to provide a more general overview of the effectiveness of acoustic and prosodic features in ASR bias discovery. The full set of extracted features was expected to exhibit substantial collinearity between frequency-based features due to the absence of vowel normalization. Table 2 presents the features contained in each of the five feature sets that were explored in the rest of this paper.

Two sets contained a single feature, namely pitch and articulation rate. These were chosen due to their suspected influence on ASR performance in Feng et al. [4]. Additionally, they are among the mentioned features in Dheram et al.'s paper on automatic cohort discovery when proposing an interpretable alternative [6].

Another two sets were inspired by the works of Adank et al. and Feng et al. The Adank set comprised vowel formants with significant regional variation between Northern Standard Dutch [17]. However, her paper measured speech characteristics of native teachers of Dutch, while the focus of the present research lies on diversity. Thus, the Feng+ set was compiled, containing vowels that were commonly misrecognized by ASR for diverse speaker groups [4]. Articulation rate and pitch were additionally added to the Feng+ set as these features were mentioned as potentially relevant sources of bias [4].

Finally, the All set was selected to imitate a scenario where one does not have extensive knowledge on existing speech variability and corresponding bias in ASR.

### 2.6. Clustering

For each feature subset, bottom-up hierarchical clustering, also called agglomerative clustering [27], was used to cluster the feature vectors into five speaker groups to match the number of known demographic groups for comparison. The algorithm starts by assigning each speaker to its own cluster, followed by repeatedly merging the closest two clusters as defined by a criterion known as a "linkage". After experimenting with different linkages using the *scikit-learn* library [24], Ward linkage [27] was found to outperform single, average and complete linkage in terms of balancing cluster sizes.

Agglomerative clustering, being a distance-based solution, was expected to better suit the task of identifying interpretable speaker groups than a density-based solution. The latter was designed to discover arbitrary cluster shapes [28]. In this paper, this was expected to complicate the comparison between speech characteristics of the resulting speaker groups, since speech variability within a cluster could fully overlap with others. Moreover, density-based solutions discard outliers, which did not align with the nature of this research as speakers with characteristics that deviate most from the mean are possibly among the most challenging for ASR systems to recognize.

Hierarchical clustering was chosen over k-means for its deterministic nature, with reproducible research in mind. While hierarchical clustering is known for its high computational cost[4], this was not a problem here due to the small sample size of 300 speakers and the maximum of 17 dimensions.

### 2.7. State-of-the-Art ASR systems

Following the setup from [5], performance and bias measures were applied to recognition output of the JASMIN data from five different ASR models. The first (*NoAug*) is a conformer model with no data augmentation on its training material. Next, *SpAug* and *SpSpecAug* are additionally trained on speed perturbed speech and the latter also on spectral augmented speech. The last two, *Whisper* and *FT-Wpr*, are OpenAI-Whisper small models [29], respectively without and with fine-tuning. All models were trained on data from the Corpus Gesproken Nederlands (CGN) [30], which consists of Dutch speech from native adults and has no speaker overlap with the JASMIN corpus.

[4] `https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering`

Table 3: *Overall Bias for each of the five ASR models under evaluation for speaker groups defined by predefined demographic metadata ("Demog.") and the clustering results of the five feature subsets ("Pitch", "Artic", "Adank", "Feng+" and "All"), for JASMIN read speech and Human-Machine Interaction separately.*

| **ASR Model** | **JASMIN Read Speech** | | | | | | **JASMIN HMI** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Demog.** | **Pitch** | **Artic** | **Adank** | **Feng+** | **All** | **Demog.** | **Pitch** | **Artic** | **Adank** | **Feng+** | **All** |
| *NoAug* | **24.6** | 4.8 | 23.1 | 11.9 | 14.8 | 12.9 | **13.1** | 5.1 | 7.8 | 10.7 | 8.6 | 9.9 |
| *SpAug* | **25.2** | 4.1 | 23.7 | 11.0 | 17.2 | 13.1 | **16.4** | 3.1 | 8.0 | 9.6 | 9.3 | 7.9 |
| *SpSpecAug* | **24.9** | 3.6 | 23.8 | 11.5 | 15.7 | 12.0 | **17.9** | 3.3 | 6.3 | 8.6 | 8.7 | 8.6 |
| *Whisper* | **21.1** | 2.9 | 21.0 | 10.9 | 14.1 | 10.9 | **18.8** | 4.9 | 12.7 | 11.7 | 12.3 | 12.5 |
| *FT-Wpr* | 24.5 | 4.6 | **24.9** | 12.4 | 15.1 | 13.5 | **13.7** | 4.7 | 7.4 | 10.2 | 7.9 | 7.6 |

The recognition output from these models was sourced from the first author of [5], and contained the insertions, deletions, substitutions and word counts of pre-processed audio files from the JASMIN corpus. Thus, while feature extraction was done on the full audio files, the corresponding ASR evaluation was carried out on a version where silence chunks were removed. This was not seen as an issue, since the speakers, speaking styles and utterances remained identical.

Note that the ASR models under evaluation were not trained by me. Instead, I made use of the insertions, deletions, substitutions and word counts of the recognition output by calculating the Word Error Rates (WERs) myself. However, the WERs of demographic groups may deviate from those in [5], as they were recalculated after excluding speakers that did not pass feature extraction.

### 2.8. Bias and Performance Measures

*Meta-measure*: The Overall Bias [31] was used to quantify the performance of the ASR models. This measure returns a single number representing how the model performs in overall, given a set of speaker groups that are being compared and a bias measure to compute the bias of a single speaker group. The Overall Bias of an ASR system is defined as

$$\text{OverallBias} = \frac{1}{G} \times \sum_{g} \text{Bias}_{spk_g} \tag{1}$$

where $G$ is the number of speaker groups minus the reference group. In this paper, the reference group is defined as the group with the lowest base error rate [1, 6]

*Bias Measure*: The bias of a speaker group was calculated as the difference between the base error rate of itself and the reference group:

$$\text{Bias}_{spk_g} = b_{spk_g} - b_{min}. \tag{2}$$

*Base Metric*: Word Error Rate (WER) was used as the base metric ($b$). The WER of a speaker group is defined as

$$\text{WER} = \frac{I + S + D}{N} \times 100\% \tag{3}$$

where $I$, $S$ and $D$ are respectively the total number of insertions, substitutions and deletions, and $N$ the total word count of all speakers in the speaker group.

### 2.9. Experimental Setup

In the experiment, clustering was applied to each feature subset and both speaking styles separately. The Word Error Rates (WERs) of the resulting speaker groups were calculated and used to estimate the bias of each ASR model using the Overall Bias measure. The average Overall Bias across the models served as an indicator of the feature sets' effectiveness in identifying speaker groups with disparities in ASR performance. For both speaking styles, the most successful feature set was further analysed. In particular, key characteristics of its resulting clusters were compared to those of demographic groups, and the demographic distribution of the clusters was examined.

## 3. Results

### 3.1. ASR Perfromance Disparities per Feature Subset

Evaluation of the Automatic Speech Recognition (ASR) models using the Overall Bias revealed that the different feature sets varied in their ability to find ASR performance disparities. Some feature sets approached the demographic baseline, while others were less effective. Table 3 presents the Overall Bias per ASR model using each feature set, for read speech and Human-Machine Interaction (HMI) separately.

For read speech, the Overall Bias resulting from each feature subset was consistent across the ASR models. The Pitch feature set was the least successful set for each model, followed by Adank, All, Feng+ and finally Artic. None of the feature sets consistently outperformed demographic groups. However, the estimated bias for the *FT-Wpr* model was higher than the baseline for the Artic feature set.

For HMI, the effectiveness of the feature sets was less consistent across the ASR models. For example, the Adank set lead to a higher Overall Bias than the Feng+ set for *FT-Wpr*, *NoAug* and *SpAug*, but not for *SpSpecAug* and *Whisper*. The Pitch set consistently performed the worst in terms of bias discovery, as was the case for read speech. The Artic set lead to the highest Overall Bias among the feature sets only for *Whisper*, while scoring the second lowest for three of the remaining ASR models. None of the feature sets outperformed the baseline on any ASR model. However, the Adank set was the most successful among the five sets of acoustic and prosodic features, despite it exclusively comprising formant frequencies.

### 3.2. Analysis of the Artic Feature Set for Read Speech

The Word Error Rate (WER) per ASR model from JASMIN read speech for each speaker group after clustering on the Artic feature set is presented in Table 4. The corresponding WERs for each demographic group can be found in Table 5.

Table 4 shows that Cluster 1 received the highest WER among the speaker groups for each ASR model, while Cluster 2 was the best recognized. The feature variation between clusters

Table 4: *%WER per ASR model for each resulting speaker group ("Cluster") when clustering on the Artic feature set, for JASMIN read speech.*

| ASR Model | Cluster | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| *NoAug* | 33.3 | **61.8** | 27.7 | 47.5 | 60.3 |
| *SpAug* | 31.6 | **62.2** | 26.4 | 46.8 | 59.6 |
| *SpSpecAug* | 30.3 | **59.2** | 23.9 | 44.3 | 56.9 |
| *Whisper* | 32.8 | **64.3** | 28.8 | 46.2 | 56.0 |
| *FT-Wpr* | 32.2 | **60.6** | 24.4 | 46.6 | 57.7 |

Table 5: *%WER per ASR model for each predefined demographic speaker group ("Group"), for JASMIN read speech.*

| ASR Model | Group | | | | |
|---|---|---|---|---|---|
| | DC | DOA | DT | NnA | NnT |
| *NoAug* | 44.5 | 29.7 | 23.9 | **62.9** | 57.0 |
| *SpAug* | 38.6 | 29.0 | 22.6 | **65.1** | 58.5 |
| *SpSpecAug* | 38.0 | 27.5 | 20.7 | **62.3** | 54.3 |
| *Whisper* | 40.3 | 34.1 | 25.5 | **58.1** | 53.8 |
| *FT-Wpr* | 40.9 | 28.2 | 22.4 | **60.8** | 57.7 |

as well as demographic groups is visualized using boxplots in Figure 1 using the Matplotlib library [32].

Due to the one-dimensional feature space, articulation rates of the clusters did not overlap. Thus Cluster 1, 4, 3, 0, and 2 were the speaker groups from lowest to highest articulation rate, respectively. This order directly corresponded with the WERs of the clusters for every ASR model, in decreasing order.

In contrast to the clusters, demographic groups generally overlapped in articulation rate. Table 5 shows that the ASR recognition was worst for nonnative speakers, NnA and NnT, among the demographic groups. While these groups had the lowest average articulation rates, DC fully overlapped in their ranges, despite the WERs being approximately 15% lower for DC than for the nonnative groups. Due to this overlap in articulation rate between demographic groups, the worst recognized cluster contained nonnative speakers as well as children. The full demographic composition of the clusters is presented in Table 6.

Table 6: *Number of speakers per demographic group that ended up in each of the resulting clusters when clustering on the Artic feature set, for JASMIN read speech.*

| Cluster | DC | DOA | DT | NnA | NnT | Total |
|---|---|---|---|---|---|---|
| **0** | 28 | 47 | 29 | 9 | 9 | 122 |
| **1** | 5 | 1 | 0 | 9 | 6 | 21 |
| **2** | 6 | 8 | 31 | 0 | 1 | 46 |
| **3** | 11 | 11 | 3 | 12 | 14 | 51 |
| **4** | 21 | 1 | 0 | 15 | 23 | 60 |

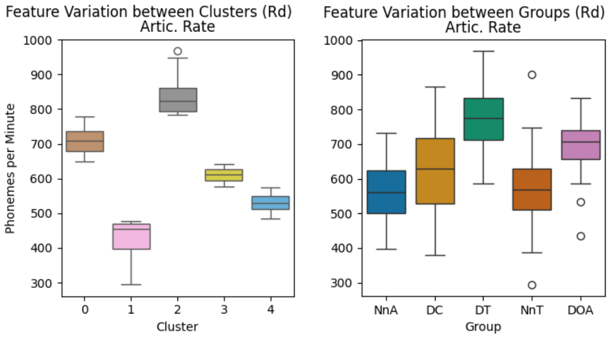

Figure 1: *Visualization of the speech variability of resulting clusters when clustering on the Artic feature set, versus predefined demographic groups, for JASMIN read speech. Each box displays the interquartile range (IQR) of the data belonging to that speaker group, corresponding to the 25th to 75th percentile, with the 50th percentile between them. The whiskers on either side of the boxes extend to points within 1.5 IQRs of the 25th or 75th percentile, and the circles represent remaining data points.*

### 3.3. Analysis of the Adank Feature Set for Human-Machine Interaction

Table 7 presents the WER per ASR model from JASMIN Human-Machine Interation (HMI) for each speaker group after clustering on the Adank feature set. The corresponding WERs for the demographic groups are shown in Table 8.

Table 7: *%WER per ASR model for each resulting speaker group ("Cluster") when clustering on the Adank feature set, for JASMIN Human-Machine Interaction (HMI).*

| ASR Model | Cluster | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| *NoAug* | 46.3 | 53.0 | 57.3 | 44.9 | **65.6** |
| *SpAug* | 43.7 | 53.2 | 55.9 | 44.5 | **59.7** |
| *SpSpecAug* | 40.6 | 47.7 | 52.0 | 40.4 | **55.6** |
| *Whisper* | 47.4 | 56.2 | **67.0** | 51.8 | 61.7 |
| *FT-Wpr* | 41.7 | 52.3 | 55.2 | 42.9 | **57.1** |

For four out of the five models, the worst ASR recognition belonged to Cluster 4. However, the WER for the *Whisper* model was highest for Cluster 2. The reference group, i.e., the group with the lowest WER, differed between models. Cluster 0 served as the reference group for three models and had the lowest average WER across the clusters.

The speech characteristics of the resulting clusters as well as demograhic groups are visualized in Figure 2. The boxplots corresponding to the resulting clusters generally showed higher average formant frequencies for Cluster 4 than for the other clusters, with an exception for F2 of /u/ and F2 of /ɔ/. When comparing the best recognized clusters 0 and 3 to the worst recognized clusters 2 and 4 in terms of ASR recognition, the pairs primarily differed in F2 of /ɑ/, F2 of /u/ and F2 of /ɔ/. However, only F2 of /ɑ/ also showed substantial variation between demographic groups.

Table 9 presents the demographic composition of the

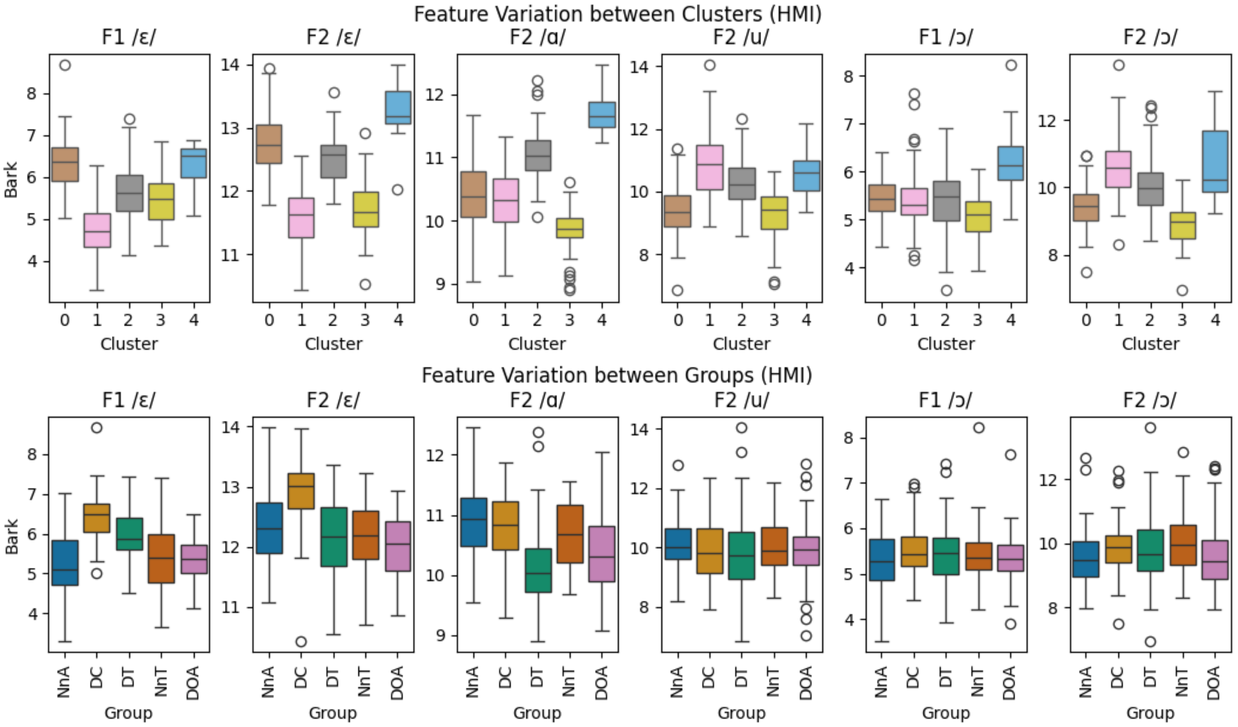

Figure 2: *Visualization of the speech variability of resulting clusters when clustering on the Adank feature set (top), versus predefined demographic groups (bottom), for JASMIN Human-Machine Interaction (HMI). Each box displays the interquartile range (IQR) of the data belonging to that speaker group, corresponsing to the 25th to 75th percentile, with the 50th percentile between them. The whiskers on either side of the boxes extend to points within 1.5 IQRs of the 25th or 75th percentile, and the circles represent remaining data points.*

Table 8: *%WER per ASR model for each predefined demographic speaker group ("Group"), for JASMIN Human-Machine Interaction (HMI).*

| ASR Model | Group | | | | |
|---|---|---|---|---|---|
| | DC | DOA | DT | NnA | NnT |
| *NoAug* | 52.0 | 43.5 | 41.7 | **63.2** | 60.8 |
| *SpAug* | 45.6 | 42.8 | 37.2 | **64.1** | 61.9 |
| *SpSpecAug* | 43.2 | 40.2 | 31.3 | **58.9** | 54.5 |
| *Whisper* | 54.5 | 50.9 | 40.6 | **73.1** | 59.3 |
| *FT-Wpr* | 43.5 | 43.9 | 37.6 | 58.3 | **59.6** |

Table 9: *Number of speakers per demographic group that ended up in each of the resulting clusters when clustering on the Adank feature set, for JASMIN Human-Machine Interaction (HMI).*

| Cluster | DC | DOA | DT | NnA | NnT | Total |
|---|---|---|---|---|---|---|
| **0** | 40 | 10 | 29 | 7 | 6 | 92 |
| **1** | 1 | 11 | 9 | 12 | 18 | 51 |
| **2** | 17 | 21 | 7 | 19 | 22 | 86 |
| **3** | 1 | 25 | 16 | 5 | 6 | 53 |
| **4** | 12 | 0 | 2 | 2 | 1 | 17 |

clusters. It shows that Cluster 0, which received the best ASR recognition, mainly comprised young native speakers. Conversely, the second best recognized cluster, Cluster 3, primarily contained native teenagers and older adults. Cluster 4, which four of the ASR models recognized the worst, predominantly consisted of native children, despite DC being the third best recognized demographic group (see Table 8). In fact, Cluster 4 lead to a higher WER for the *NoAug* model than the worst recognized demographic group NnT, consisting of nonnative adults.

## 4. Responsible Research

In any form of research, it is crucial to consider reproducibility and ethical implications. In this section, the reproducibility of the results, inclusion of the paper, and ethical implications of the proposed approach are addressed.

### 4.1. Reproducibility

Due to space limitations, only the most significant results are presented in this paper. Given the broad scope of the proposed approach in this paper, it was infeasible to further analyse the performances of individual clusters resulting from the remaining feature sets. However, results can be reproduced as the full implementation is publicly available on GitHub[5]. The extracted features are not shared to prevent any reverse engineering of the JASMIN-CGN corpus. The corpus is

[5] https://github.com/Kayyleigh/Interpretable-Automatic-Bias-Discovery-in-ASR

available on request from the Dutch Language Institute[6] for research purposes.

### 4.2. Inclusion

Inclusion is achieved in this paper through the use of colorblind-friendly plots using the Seaborn Python visualization library [33].

### 4.3. Ethical Considerations

The aim of this research was not to predict demographic labels of unknown speakers. While ideas were borrowed from earlier works that measure the fluency of native and nonnative speakers [14, 15, 16], these papers mention definitions of fluency that do not align well with the focus of this research. The goal of fair ASR is to ensure all speakers are recognized equally regardless of demographics. This reflects the fact that nonnative speech is generally comprehensible by speakers of the Dutch language, evidenced by the ability to transcribe their speech. Thus, it is important to highlight that the performance measures in this research exclusively describe the quality of the ASR systems.

## 5. Discussion

This study explored the effectiveness of clustering acoustic and prosodic features in the discovery of performance disparities in Automatic Speech Recognition (ASR). In particular, five feature sets were compiled and each clustered into speaker groups, which were then evaluated in terms of bias and compared to known bias between demographic groups. The experiment was carried out for read speech and Human-Machine Interaction (HMI) separately. The results suggest that certain feature sets can discover substantial performance disparities for different ASR models. However, performance varies across feature sets, and no relation seems to exist between the size of a set and its performance.

The five sets were motivated by different parts of the literature. Two single-feature sets, one comprising Pitch and the other Articulation Rate, were chosen as these are expected to vary between age groups and levels of nativeness [4]. Next, two sets were compiled based on the works of Adank [17] and Feng [4], respectively. The remaining feature set contained all extracted features.

The Pitch set lead to the smallest disparities for both speaking styles, which suggests pitch may be less influential for bias than expected. However, note that pitch is primarily known to vary across ages, while the highest WERs for JASMIN data belonged to nonnative speakers, of which no young children were present in the data.

Conversely, the Artic set lead to substantial bias for read speech for each ASR model. In fact, the *FT-Wpr* model received a higher Overall Bias when using the clusters than when using demographic groups. When visualizing the articulation rates, the Word Error Rate (WER) of resulting clusters were found to decrease as the articulation rates increased (see Figure 1). Known ASR performance for demographic groups [5] were highest for nonnative groups. The findings are in line with this, as these groups exhibited the lowest average articulation rate among demographic groups. However, native children showed substantial variation in articulation rate, in line with findings from Lee et al. [12]. As a result, children were spread across clusters of varying ASR performance, including the worst recognized cluster. This suggests that more meaningful disparities may be discovered between ages when further splitting the DC speaker group.

A surprising finding was that the Artic set generally performed worse than the other sets for HMI. Instead, the most successful features set was Adank, exclusively comprising formant frequency features. Interestingly, the highest WER resulting from this feature set belonged to a cluster that primarily consisted of native children. This contrasts with previous research [5], where ASR systems consistently recognized nonnative speaker groups more poorly than native children. Moreover, one of the best recognized clusters, Cluster 0, mainly comprised young native speakers as well. This further suggests that significant ASR performance disparities exist within demographic groups.

The findings from successful feature sets for read speech and HMI collectively imply that clustering on acoustic and prosodic features can lead to poorly recognized groups of speakers that do not necessarily align with demographics. However, for the two feature sets discussed in this paper, the clusters with the highest WERs were also the smallest clusters. Furthermore, considerable variation could be observed between cluster sizes in general, and no reference clusters were identified with a lower WER than the reference demographic group DT.

The main benefit of using acoustic and prosodic features instead of Machine Learning-based speaker embeddings is that the key characteristics of poorly recognized speaker groups can be better interpreted. However, note that the ability of a feature set to discover ASR performance disparities does not necessarily mean the features cause bias; they may instead be an effective proxy for demographic information. Nevertheless, the proposed approach enables interpretable ASR bias discovery without the need for demographic metadata.

### 5.1. Limitations

The proposed approach has several limitations. First and foremost, formant frequencies were extracted using the same Praat configuration for each speaker, despite this being discouraged due to anatomical differences between genders and age groups. Consequently, the accuracy of these measurements may vary between speakers. Additionally, the present approach still requires an expected number of clusters. When the approach is applied for bias discovery without a direct comparison to known disparities between demographic groups, the usefulness of the resulting clusters is presumably dependent on an expert's interpretation, which is unknown beforehand. Moreover, this interpreted meaning of the resulting clusters is likely to differ across datasets and speaking styles. Finally, no quantitative assessment of the interpretability of resulting bias was carried out due to time constraints.

A limitation of the study is that the results only hold for the five ASR models under evaluation. Additionally, the feature sets in this paper were exclusively designed for the Dutch language. Finally, the approach may benefit from vowel normalization [9], which was not applied in this study due to multiple vowels being absent in a considerable fraction of HMI data.

### 5.2. Future Research

Further directions for research include investigating how extracted features from speakers with anatomical differences can be obtained and compared more fairly in terms of

[6] `https://taalmaterialen.ivdnt.org/download/tstc-jasmin-spraakcorpus/`

configurations as well as vowel normalization techniques. Furthermore, the approach can be extended with a quantitative assessment of cluster interpretability. The methodology may be improved by adding features that capture rhythm and vowel space [34], as well as additional prosodic features [14, 15] and word choice. Finally, future directions include comparing alternative scaling methods and clustering algorithms, as well as implementing a systematic approach for feature subset selection.

## 6. Conclusions

This study investigated the effectiveness of acoustic and prosodic features in discovering interpretable performance disparities Automatic Speech Recognition (ASR) without the need for demographic metadata. Five linguistically-motivated feature sets were compared. The study shows that certain feature sets successfully lead to clusters with varying ASR recognition. In particular, when clustering exclusively on articulation rate, larger disparities are found than for demographic groups for read speech. For Human-Machine Interaction (HMI), none of the sets outperformed the demographic baseline. However, the most successful of the five feature sets comprised only formant frequencies. For both speaking styles, resulting clusters of the best feature sets consisted of diverse demographics, indicating a potential to find new ASR performance disparities that may not exist between demographic groups. Interestingly, an exception was the worst-recognized cluster for HMI, which primarily comprised children despite previous studies finding that the nonnative speaker groups exhibit larger Word Error Rates (WERs) than the native groups.

This paper provides a first report on interpretable fairness in ASR without the need for demographic metadata. Future research should focus on improving the approach by adding new features and exploring alternative methods. By doing so, an improved approach may be found that, given an appropriate language-specific feature space, has the potential to outperform demographic groups in the discovery of bias in ASR systems for any spoken language.

## 7. Acknowledgements

I would like to express my deepest gratitude to Dr. Odette Scharenborg and Dr. Jorge Martinez Castaneda for their continuous support and guidance. Special thanks should also go to YuanYuan Zhang and Dr. Tanvina Patel for their assistance in accessing the data. Lastly, I would like to thank Fanni Fiedrich, Vincent van Vliet, and my two anonymous peer reviewers for their insightful feedback on the drafts.

## 8. References

[1] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, “Quantifying Bias in Automatic Speech Recognition,” *arXiv preprint arXiv:2103.15122*, Apr. 2021.

[2] M. K. Ngueajio and G. Washington, “Hey ASR system! why aren’t you more inclusive? Automatic Speech Recognition systems’ bias and proposed bias mitigation techniques. a literature review,” 2022, vol. 13518, pp. 421–440.

[3] M. Fuckner, S. Horsman, P. Wiggers, and I. Janssen, “Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers,” in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2023, pp. 146–151.

[4] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, “Towards inclusive automatic speech recognition,” *Computer Speech Language*, vol. 84, p. 101567, 2024.

[5] T. Patel, W. Hutiri, A. Ding, and O. Scharenborg, “How to evaluate automatic speech recognition: Comparing different performance and bias measures,” unpublished.

[6] P. Dheram, M. Ramakrishnan, A. Raju, I.-F. Chen, B. King, K. Powell, M. Saboowala, K. Shetty, and A. Stolcke, “Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities,” *Interspeech 2022*, Sep. 2022.

[7] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” *arXiv preprint arXiv:1710.10467*, 2017.

[8] P. Ladefoged and D. E. Broadbent, “Information Conveyed by Vowels,” *The Journal of the Acoustical Society of America*, vol. 29, no. 1, pp. 98–104, Jan. 1957.

[9] P. Adank, “Vowel normalization: a perceptual-acoustic study of Dutch vowels,” Ph.D. dissertation, Catholic University of Nijmegen, 2003.

[10] A. Potamianos, S. Narayanan, and S. Lee, “Automatic speech recognition for children,” 1997, pp. 2371–2374.

[11] C. Cucchiarini, H. Van hamme, O. van Herwijnen, and F. Smits, “JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*. Genoa, Italy: European Language Resources Association (ELRA).

[12] S. Lee, A. Potamianos, and S. Narayanan, “Analysis of children’s speech: duration, pitch and formants,” in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997, pp. 473–476.

[13] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, “Effects of aging on vocal fundamental frequency and vowel formants in men and women,” *Journal of voice : official journal of the Voice Foundation*, vol. 32, no. 5, pp. 644.e1–644.e9, Sep. 2018.

[14] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology,” *The Journal of the Acoustical Society of America*, vol. 107, pp. 989–99, Mar. 2000.

[15] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech,” *The Journal of the Acoustical Society of America*, vol. 111, pp. 2862–73, Jul. 2002.

[16] C. Cucchiarini, J. van Doremalen, and H. Strik, “Fluency in non-native read and spontaneous speech,” in *Proc. DiSS-LPSS Joint Workshop (DiSS 2010)*, 2010, pp. 15–18.

[17] P. Adank, R. van Hout, and H. van de Velde, “An acoustic description of the vowels of Northern and Southern standard Dutch II: Regional varieties,” *J. Acoust. Soc. Am.*, vol. 121, no. 2, 2007.

[18] D. Crystal, *A dictionary of linguistics and phonetics*, 6th ed., ser. The language library. Oxford: Blackwell, 2008.

[19] G. Fant, *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations*, ser. Description and analysis of contemporary standard Russian. The Hague: Mouton, 1960.

[20] P. Adank, R. Van Hout, and R. Smits, “An acoustic description of the vowels of Northern and Southern Standard Dutch,” *The Journal of the Acoustical Society of America*, vol. 116, no. 3, pp. 1729–1738, Sep. 2004.

[21] P. Boersma and D. Weenink, “Praat: doing Phonetics by Computer,” 2024. [Online]. Available: https://www.fon.hum.uva.nl/praat/

[22] L. Robles and M. A. Ruggero, "Mechanics of the Mammalian Cochlea," *Physiological Reviews*, vol. 81, no. 3, pp. 1305–1352, Jul. 2001.

[23] B. M. Lobanov, "Classification of Russian Vowels Spoken by Different Speakers," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 606–608, Feb. 1971.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*.

[25] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[26] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[27] J. H. W. Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[28] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise."

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022.

[30] N. Oostdijk, "The spoken Dutch corpus. overview and first evaluation," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer, Eds. Athens, Greece: European Language Resources Association (ELRA), May 2000.

[31] T. Patel and O. Scharenborg, "Using Data Augmentations and VTLN to Reduce Bias in Dutch End-to-End Speech Recognition Systems," Jul. 2023.

[32] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[33] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[34] L.-F. Lai, J. G. V. Hell, and J. Lipski, "The Role of Rhythm and Vowel Space in Speech Recognition," in *Speech Prosody 2022*. ISCA, May 2022, pp. 425–429.