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

Automatic Speech Recognition (ASR) systems misrecognize certain people more than others [1]. To measure these disparities, we need demographic information of the speakers, which is often unavailable. Recent work uses machine learning to cluster similar voices [2], but this black-box approach leads to uninterpretable speaker groups. What if we cluster the voices by interpretable attributes instead?

Research Question

> How effective are language-specific **acoustic and prosodic feature sets** in the discovery of **interpretable** Automatic Speech Recognition (ASR) **performance disparities** between groups of similar sounding speakers?

Author

Kayleigh Jones
✉ k.m.jones@student.tudelft.nl

Supervisors

Dr. Odette Scharenborg
Dr. Jorge Martinez Castaneda

References

[1] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, 'Towards inclusive automatic speech recognition', Comput. Speech Lang., vol. 84, p. 101567, Mar. 2024, doi: 10.1016/j.csl.2023.101567.
[2] P. Dheram et al., 'Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities', presented at the Proc. Interspeech 2022, 2022, pp. 1268–1272. doi: 10.21437/Interspeech.2022-10816.
[3] Cucchiarini, C., van Herwijnen, O., Smits, F., et al., 2006. JASMIN-CGN: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In: Proc. LREC.
[4]T. Patel, W. Hutiri, A. Ding, and O. Scharenborg, "How to evaluate automatic speech recognition: Comparing different performance and bias measures," unpublished.
[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022.
[6] 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.
[7] P. Boersma and D. Weenink, 'Praat: doing Phonetics by Computer'. Accessed: May 10, 2024. [Online]. Available: https://www.fon.hum.uva.nl/praat/
[8] T. Patel and O. Scharenborg, "Using Data Augmentations and VTLN to Reduce Bias in Dutch End-to-End Speech Recognition Systems," Jul. 2023.

TU Delft
Delft University of Technology

Materials

- JASMIN-CGN [3]: Database of diverse Dutch speech.

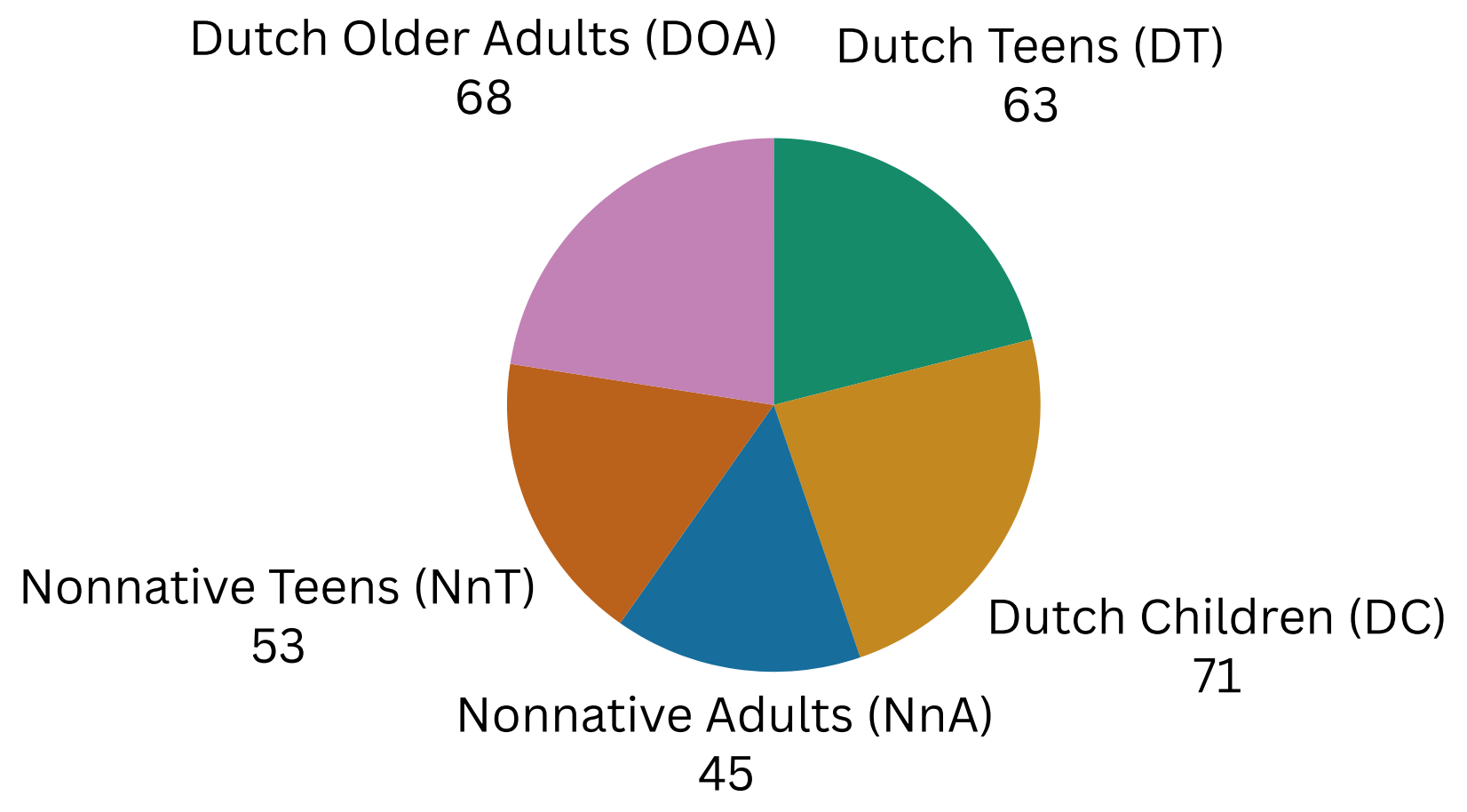

Figure 1: *#Speakers per speaker group in JASMIN, region NL.*

- Recognition output for the same JASMIN data of five pre-trained ASR models [4]: *NoAug*, *SpAug*, *SpSpecAug*, *Whisper*, and *FT-Wpr*. The first three are conformer models, the rest are OpenAI-Whisper small models [5].

Methodology

- Five feature sets based on research from Adank et al. [6] and Feng et al. [1], extracted using Praat [7].
- Min-max scaling
- Agglomerative Clustering with Ward linkage
- Evaluation of ASR models using Overall Bias [8].

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	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: *feature sets to cluster on. All features were averaged per speaker. Pitch was only measured over phoneme segments.*

Performance metrics

Overall bias of an ASR model [8], for *G* groups (excluding the *min* group) :

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

Bias of a group, *min being* the group with the best base performance:

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

The Word Error Rate (WER): base metric (*b*) of a speaker group:

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

Results

- No feature set consistently outperformed demographics in finding ASR performance disparities
- The Pitch set lead to the lowest found Overall Bias for both speaking styles
- For read speech, Artic outperformed demographic groups for a single ASR model
- For Human-Machine Interaction (HMI), Adank performed best out of the five sets, while Artic performed second worst, despite Adank comprising only formant features.

The Artic set for read speech and the Adank set for HMI were further analysed.

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

Table 2: *Overall Bias per model, with demographic groups vs the feature sets, for JASMIN read speech and HMI.*

Artic - read speech

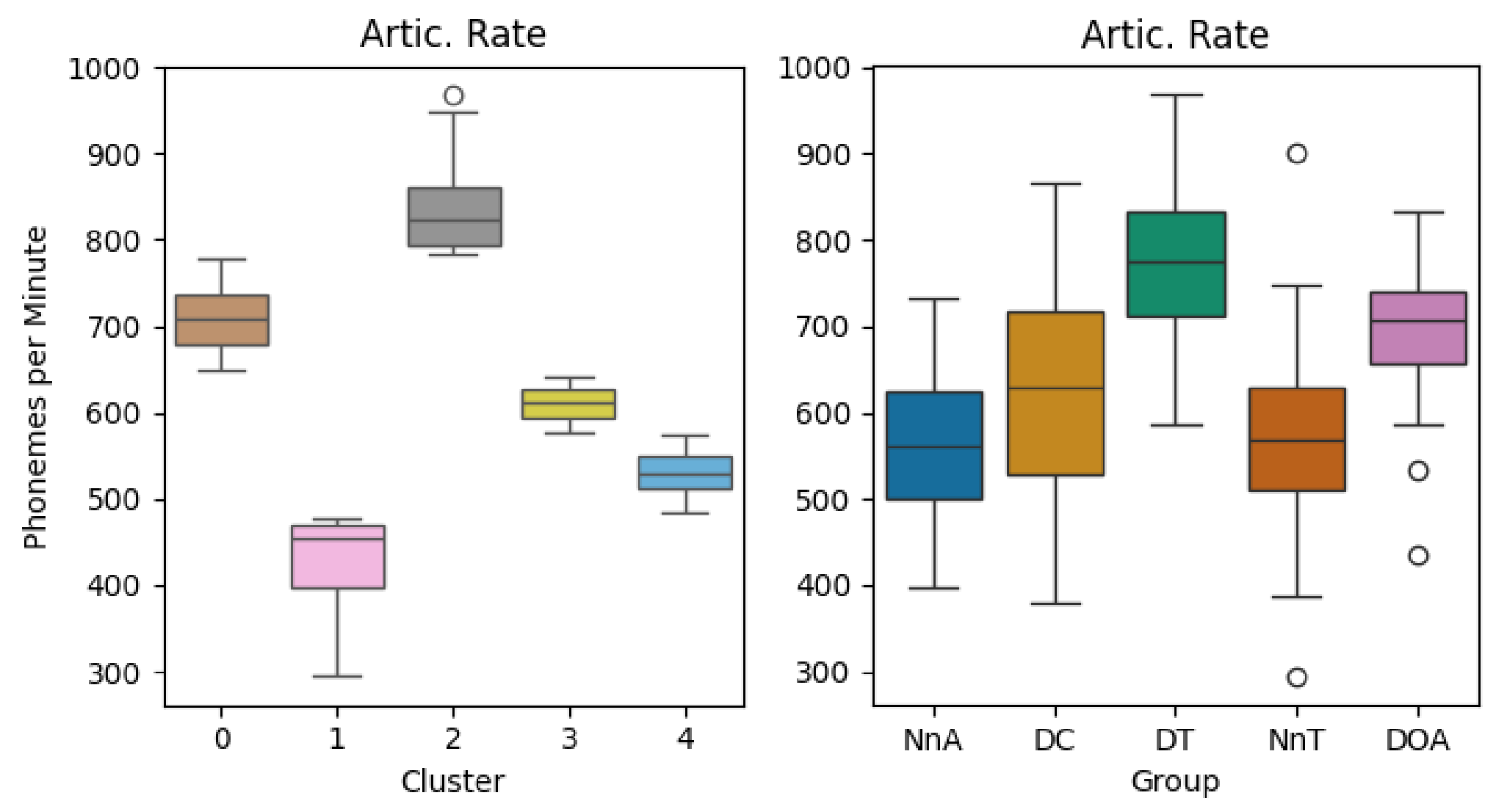

Figure 2: *Artic. rate variability of clusters (left) vs demographic groups (right), for JASMIN read speech.*

Notable findings

- Clusters with the fastest to slowest articulation rate got the best to worst ASR recognition, respectively
- Clusters with the highest WERs (1 and 4) contained nonnatives (71.4% and 63.3%) as well as Dutch children (23.8% and 35%)

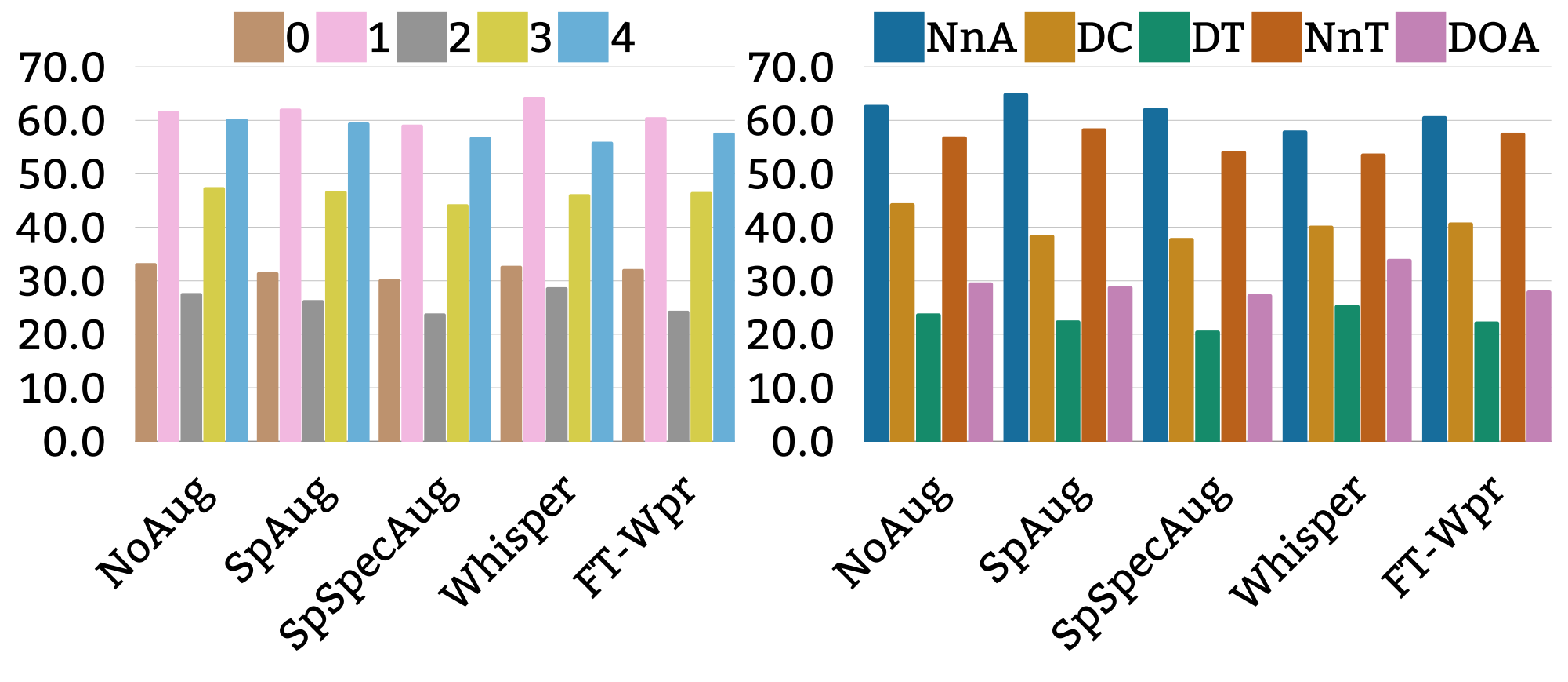

Figure 3: *Word Error Rate (WER) of clusters on the Artic set (left) or demographic groups (right) per ASR model, for JASMIN read speech.*

Adank - HMI

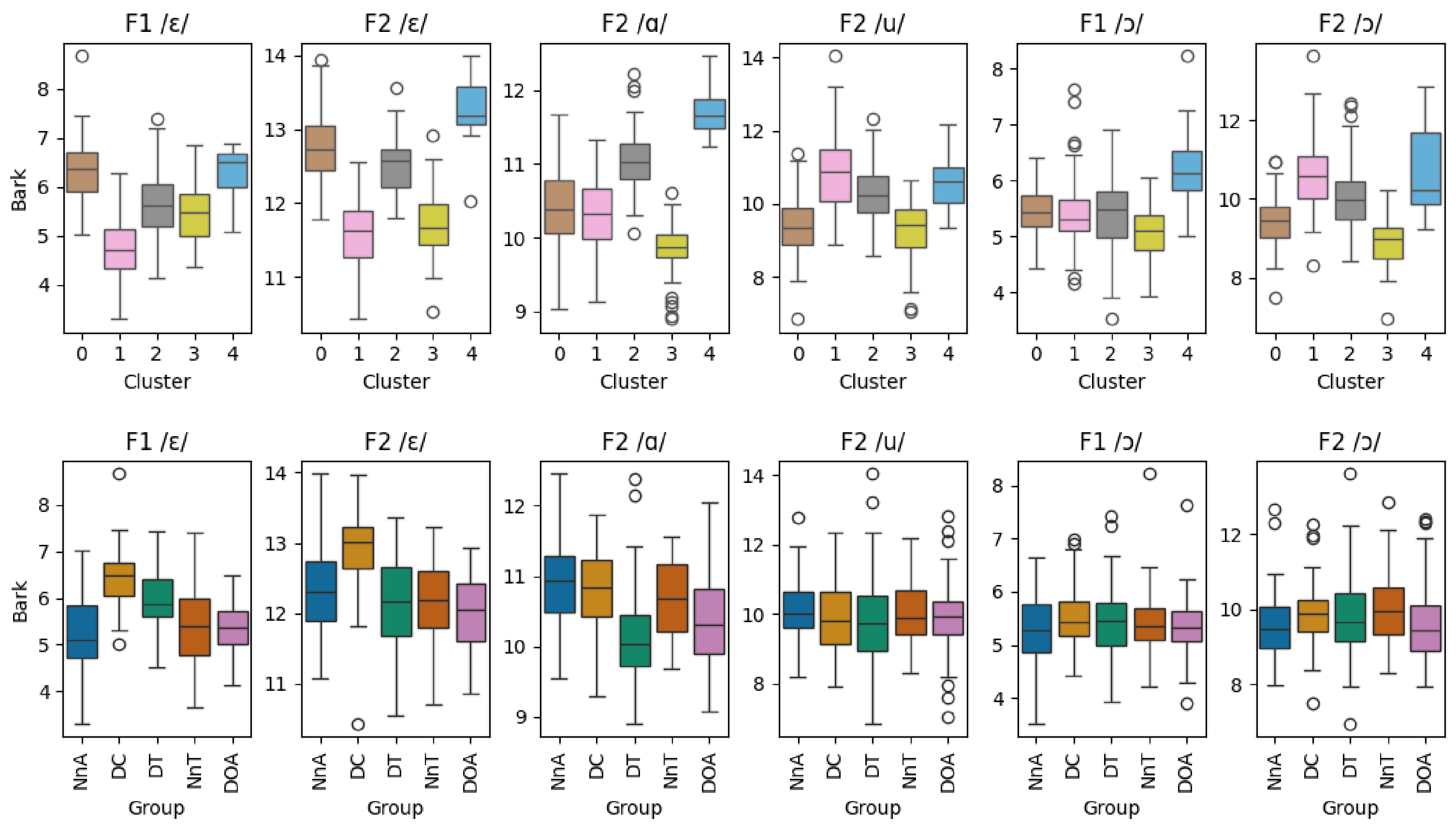

Figure 4: *Speech variability of clusters (top) on the Adank set vs demographic groups (bottom) for JASMIN HMI.*

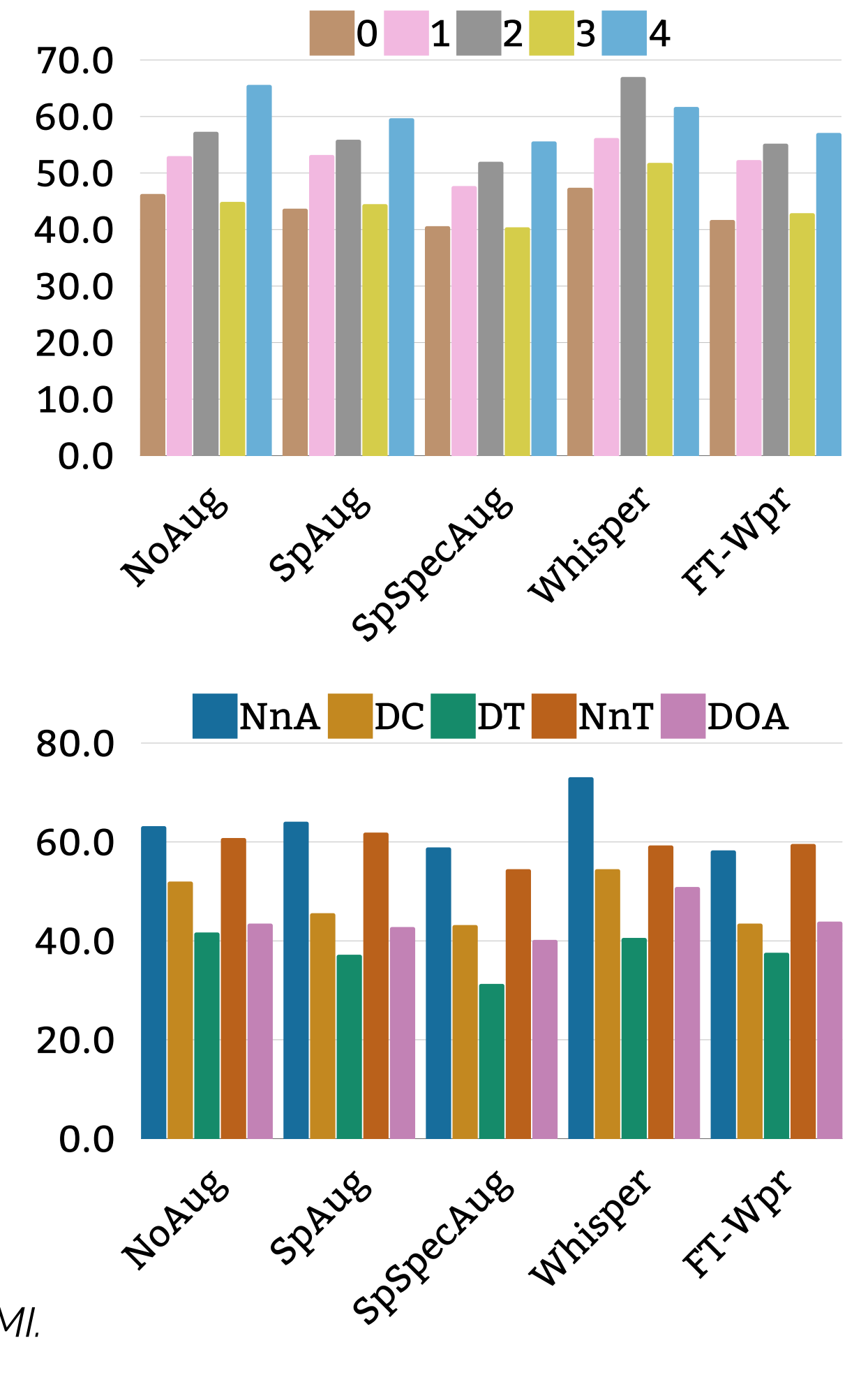

Figure 5: *WER per cluster on the Adank set (top) or demographic group (bottom) per ASR model, for JASMIN HMI.*

Notable findings

- The *NoAug* model performed worse on Cluster 4 than on the demographic group NnA, while 70.6% of Cluster 4 comprised Dutch children

Limitations

- Praat configuration ideally depends on gender/age. Default (= for adult women) was used, but the fact that it matters does not align with the goal of this research
- In practice, there is no optimal #clusters since there is no ground truth. Interpretation requires expert knowledge: different data/algorithms/ #clusters lead to groups with different speech variability
- Due to time constraints, cluster interpretability could not be assessed quantitatively

Discussion and Conclusions

Acoustic and prosodic feature sets uncovered ASR biases similar to ones between demographic groups. Effectiveness differed between speaking styles, and not all sets lead to significant disparities.

The most successful sets showed that significant biases can occur within demographic groups. The proposed approach has the potential to uncover these such that key speech characteristics of resulting speaker groups can be compared.

However, none of the sets consistently outperformed the demographic approach. Future work should experiment with alternative pre-processing and clustering methods.