AfriBED: a Speech Recognition Bias Evaluation Dataset for Low Resource African Languages

Claytone Sikasote
Department of Computer Science
University of Cape Town
Cape Town, South Africa
skscla001@myuct.ac.za

Hussein Suleman
Department of Computer Science
University of Cape Town
Cape Town, South Africa
hussein@cs.uct.ac.za

Jan Buys
Department of Computer Science
University of Cape Town
Cape Town, South Africa
jan.buys@uct.ac.za

***Abstract*—Despite advancements in automatic speech recognition systems (ASR), current systems exhibit speaker-attribute bias, i.e., their speech recognition performance is biased against certain speakers based on their demographic attributes such as gender, accents, and speaker type. However, accurately assessing speaker attribute bias in a given ASR system requires the availability of a bias evaluation dataset with comparable speech across subgroups of a speaker attribute. In this paper, we introduce African speech Bias Evaluation Dataset (AfriBED), a novel speech recognition bias evaluation dataset for two low resource African languages, Bemba and Nyanja, to detect and estimate gender and speaker-type bias for ASR systems. The dataset comprises 9 hours of speech with 3928 utterances recorded by 80 speakers for Bemba, while the Nyanja set includes 6260 utterances with approximately 18 hours of speech recorded by 127 speakers. The results of our baseline experiments demonstrate that the datasets are useful for examining gender and speaker-type attribute bias in the target languages.**[1]

***Index Terms*—ASR, Fairness, Bemba, Nyanja, Africa languages**

I. Introduction

In recent years, there has been a growing interest in developing ASR systems for low resource African languages [1]. With the majority of languages spoken across the continent classified as low resource languages (LRLs), most research efforts have focused on building training data dataset in order to improve ASR accuracy [2]–[5]. Although this research address the inherent limitations that characterize LRLs, one research dimension that has not received significant attention is how to effectively evaluate speech recognition models for speaker attribute biases [6].

In this paper, we present our data collection effort to address the issue of limited speech recognition bias evaluation datasets for African languages [7], [8]. We introduce a novel bias evaluation dataset containing read speech in two low resource African languages: Bemba and Nyanja. The dataset is created for evaluating ASR systems for gender and speaker-type (native and non-native speaker) biases. We carefully constructed the datasets to ensure that the subgroups for each gender and speaker-type attribute have the same speech content to allow fair performance comparison. This is achieved by recording the same read text sentences by speakers of each speaker attribute subtype. To the best of our knowledge, our work is the first to create dedicated ASR bias evaluation datasets to examine speaker-attribute bias in ASR systems for Bemba and Nyanja.

[1]The dataset is publicly available at: https://github.com/csikasote/AfriBED

II. Related work

Several bias evaluation datasets exist to evaluate ASR systems for various forms of speaker-attribute bias [9]–[13]. The most popular is the Artie Bias Corpus [10] based on the Mozilla Common Voice corpus [14] for age, gender, and accent bias detection. Liu et al. [12] created the Casual Conversations dataset for gender, age, and skin tone. Shriberg et al. [15] curated 72 hours of conversation-like speech based on dialogues from naturally-occuring meetings. There are others like the recently introduced FairSpeech [11] dataset. However, the majority of these datasets are designed for English ASR systems, leaving the majority of languages without tools to examine speaker attribute bias, particularly LRLs. In addition, the speech across attribute subgroups in these datasets is not the same. The subgroups of attributes do not have speech drawn from the same text sentences, resulting in unfair comparisons. In contrast, our study aims to create bias evaluation datasets for LRLs, targeting African languages. Therefore, we create a novel speech recognition bias evaluation dataset that covers two Sub-Saharan African languages: Bemba and Nyanja, which are spoken predominantly and widely in Zambia and neighboring countries [16].

III. Methodology

In this section, we provide details of the different phases of our data collection process. We highlight that we obtained ethical clearance for data collection and followed all ethical standards for responsible data collection practices.

A. Demographic survey

We started by conducting a demographic survey among students at the University of Zambia, targeting students in the Department of Computing and Informatics using convenient sampling for our selection criteria. The survey was carried out for three reasons: to determine the native languages (amongst languages spoken in Zambia) spoken among the students who are our target population; to estimate the number of students we could invite to participate voluntarily in the successive

phases; and to establish the demographic information that we could consider for our bias evaluation dataset. The demographic survey was set up using the Google Forms platform and sent out to the students' university email addresses.

A total of 359 students responded to the survey. The responses showed that Bemba (41.23%) and Nyanja (44.57%) were the most populous native languages spoken among students and spoken as their first or second language. None of the other languages spoken had a critical mass. Gender and speaker-type i.e., native and non-native speakers, were the two possible speaker attributes that could be captured for our data collection. [2] We did not find sufficiently large subgroups for other attributes such as race, age and region. All participants identified themselves as black, belonging to a single age group, between 18 and 25 years, and the majority (more than 70%) came from one region, Lusaka.

B. Generating text sentences

Due to the lack of high-quality public textual data in our target languages, we obtained 1000 English text sentences from the test split of the open-source ASR dataset, LibriSpeech [17] and translated them into Bemba and Nyanja for audio recording. The LibriSpeech dataset comprises 1000 hours of speech in the English language derived from publicly domain audio books from the LibriVox project. [3] To ensure that we obtained high-quality text translations, we recruited two paid professional translators, one translator per language, to translate the English sentences into the target languages. In addition, for each language, we recruited three native speakers (students) on a voluntary basis to verify and ensure that no mismatch occurred in the English-to-translation sentence pairs.

C. Creating audio recordings

We recruited a total of 205 students to record speech from translated text sentences on a voluntary basis. Most of the students were drawn from the participants in the survey who expressed interest in participating in the audio recording task. We carefully designed our recording strategy so that the speech between subgroups of the speaker attributes (gender and speaker type) is the same to ensure subgroup comparability. To achieve this, we defined two sets of speaker attribute groups: *primary groups* (*native male*, *non-native male*, *native female*, and *non-native female*) and *secondary groups* (*male*, *female*, *native* and *non-native*). The primary groups are intersectional attribute groups between gender and speaker-type attribute groups, while secondary groups are direct subgroups of gender (male and female) and speaker-type (native and non-native) attributes, whose data composition can be worked out and aggregated from the corresponding primary groups. To assign text sentences for audio recording to participants, we grouped the participants according to their self-identified primary groups. We then split the 1000 text sentences into two batches: *Batch A*, containing sentences 1–500, and *Batch B*, containing sentences 501–1000, and assigned them for recording to each primary group as shown in Fig. 1. Furthermore, we split the text sentences in each batch into portions of 50 text sentences, which we then assigned to each participant to record. To ensure speech variation and diversity, and to allow in-group inter-speaker evaluation, we assigned each 50 sentence block to at least three participants. Using this strategy, gender and speaker-type subgroups can be assigned from the primary groups with speech in the subgroups based on the same sentences and without speaker overlap. The audio files were recorded using the Audacity software, with a sampling rate of 16kHz and 32-bit float PCM sample encoding setting. [4] All the utterances are within 20 seconds.

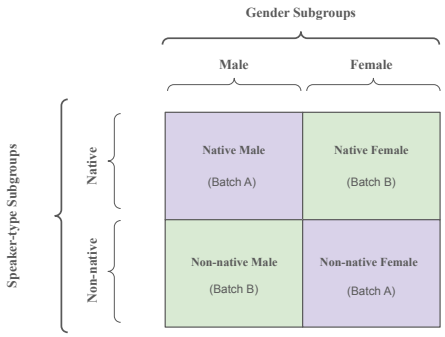

Fig. 1. Our strategy for allocating text sentences to primary subgroups to create audio recordings, ensuring comparable speech between subgroups of both gender and speaker-type attributes and without speaker overlap.

TABLE I
SUMMARY STATISTICS OF THE BIAS EVALUATION DATASETS WITH NATIVE MALE (NM), NATIVE FEMALE (NF), NON-NATIVE MALE (NNM) AND NON-NATIVE FEMALE (NNF) PRIMARY INTERSECTIONAL SPEAKER ATTRIBUTE SUBGROUPS.

Subgroup	Bemba			Nyanja		
	Files	Speakers	Hours	Files	Speakers	Hours
Overall sets						
NM	1464	30	3.19	2673	54	7.04
NF	821	17	1.86	1144	23	3.12
NNM	1195	24	3.28	1894	39	6.85
NNF	448	9	0.99	549	11	1.52
Total	**3928**	**80**	**9.32**	**6260**	**127**	**18.53**
Mini-version sets						
NM	350	7	0.79	299	6	1.10
NF	349	7	0.73	299	6	0.83
NNM	348	7	0.87	300	6	0.84
NNF	349	7	0.74	300	6	0.80
Total	**1396**	**28**	**3.13**	**1198**	**24**	**3.57**

[2]We define the first language as the native language the speaker has been exposed to since birth. A native speaker is one where the language is their first language, while a non-native speaker is one where the language is not their first language but they declare a good degree of proficiency in reading, understanding and writing.

[3]https://librivox.org/

[4]https://www.audacityteam.org/

TABLE II

RESULTS OF FINE-TUNED ASR MODELS EVALUATED ON OVERALL SETS OF THE LANGUAGE DATASETS. WE REPORT CER ON THE FOLLOWING SUBGROUPS: NATIVE MALE (N), NON-NATIVE MALE (NNM), NATIVE FEMALE (NF), NON-NATIVE FEMALE (NNF), MALE (M), FEMALE (F), NATIVE (N), NON-NATIVE (NN) AND COMBINED (C). RESULTS ARE AVERAGED OVER THREE RUNS. LOWER CER AND CER-GAP VALUES ARE BETTER. WE REPORT MEAN AND STANDARD DEVIATION (IN PARENTHESES) VALUES.

Language	Model	Male			Female			Gender				Speaker-type			
		NM	NNM	CER-gap	NF	NNF	CER-gap	M	F	C	CER-gap	N	NN	C	CER-gap
Bemba	SSA-HuBERT-L-60K	16.76 (0.17)	19.38 (0.13)	**2.62 (0.06)**	17.43 (0.38)	21.09 (0.45)	**3.66 (0.24)**	17.95 (0.15)	18.71 (0.39)	18.19 (0.19)	0.77 (0.35)	17.00 (0.21)	19.84 (0.17)	18.19 (0.18)	**2.84 (0.10)**
	SSA-HuBERT-XL-60K	15.45 (0.09)	18.29 (0.13)	2.84 (0.05)	15.69 (0.34)	19.85 (0.17)	4.17 (0.21)	16.74 (0.10)	17.15 (0.27)	16.87 (0.16)	**0.41 (0.18)**	15.53 (0.17)	18.70 (0.13)	16.87 (0.16)	3.17 (0.05)
	MMS-300M	20.21 (0.10)	21.81 (0.02)	**1.60 (0.09)**	21.30 (0.81)	23.41 (0.20)	**2.11 (0.63)**	20.94 (0.06)	22.04 (0.59)	21.29 (0.23)	1.11 (0.53)	20.61 (0.35)	22.24 (0.07)	21.29 (0.24)	**1.63 (0.30)**
	MMS-1B-All	14.34 (0.18)	17.55 (0.08)	3.21 (0.12)	14.40 (0.10)	18.54 (0.06)	4.14 (0.07)	15.80 (0.13)	15.85 (0.08)	15.82 (0.11)	**0.06 (0.08)**	14.36 (0.13)	17.82 (0.08)	15.82 (0.11)	3.45 (0.07)
Nyanja	SSA-HuBERT-L-60K	25.77 (0.19)	56.09 (1.02)	30.32 (0.88)	25.81 (0.05)	40.91 (0.88)	15.10 (0.93)	38.58 (0.52)	30.54 (0.24)	36.39 (0.43)	**8.04 (0.34)**	25.78 (0.12)	52.77 (0.98)	36.39 (0.44)	26.99 (0.89)
	SSA-HuBERT-XL-60K	24.89 (0.14)	54.62 (0.68)	**29.73 (0.74)**	24.72 (0.26)	39.72 (0.95)	**14.10 (0.94)**	37.46 (0.27)	29.42 (0.37)	35.27 (0.30)	**8.04 (0.13)**	24.84 (0.17)	51.37 (0.74)	35.27 (0.30)	**26.53 (0.78)**
	MMS-300M	29.35 (1.12)	49.28 (1.52)	**19.94 (2.56)**	28.53 (0.66)	42.61 (0.60)	**14.09 (0.64)**	37.78 (0.32)	32.95 (0.57)	36.46 (0.32)	**4.83 (0.54)**	29.09 (1.00)	47.83 (1.06)	36.46 (0.32)	**18.74 (1.95)**
	MMS-1B-All	22.35 (0.08)	55.63 (0.24)	33.27 (0.20)	21.19 (0.09)	38.80 (0.29)	17.60 (0.21)	36.41 (0.13)	26.72 (0.15)	33.76 (0.14)	9.69 (0.04)	21.99 (0.08)	51.95 (0.25)	33.76 (0.14)	29.96 (0.19)

D. Data validation

We validated all data collected following the strategy of Conneau et al. [18] to ensure that only clean and high quality audio recordings were retained. All valid recordings, from either quiet or noisy environments, are kept as is.

E. Resulting Datasets

Table I provides summary counts of audio files, number of speakers, and size (in hours) for each subgroup in the resulting evaluation sets. The overall datasets comprise a total of 9.32 hours of Bemba speech (3,928 audio files from 80 speakers) and 18.53 hours of Nyanja speech (6,260 audio files from 127 speakers). In addition, we curated subsets of the overall datasets into mini-version datasets, which include only speech recordings for text segments with at least one contribution from each subgroup, ensuring balanced and comparable evaluation across subgroups. The mini-version datasets include 3.13 hours of Bemba speech (1,396 files from 28 speakers) and 3.57 hours of Nyanja speech (1,198 files from 24 speakers). The dataset, along with transcriptions and anonymized metadata files, are released publicly to the research community under a Creative Commons BY-NC license.[5]

IV. BASELINE EXPERIMENTS

To provide baseline speech recogntion and fairness performances of speech models in the target languages, we fine-tune two popular open-source pre-trained speech models: massively multilingual model for speech (MMS) [19] and the Africa-centric self-supervised multilingual speech model for Sub-Sahara African languages (SSA-HuBERT-Large-60K) [20] on existing open-source labeled ASR datasets for Bemba (28 hours) and Nyanja (25 hours) from Zambezi Voice [3]. The datasets are not balanced with speaker-type and gender attribute speech. We use the 300M and 1B parameter variants of both models. We report character error rate (CER) metric [21] for ASR performance. For bias, we use the absolute difference in ASR performance between the subgroups of the respective speaker attributes (CER-gap) [22]. We use our full collected dataset to evaluate the fine-tuned speech models.

[5]The dataset is publicly available at: https://github.com/csikasote/AfriBED

V. RESULTS

Table II provides the results of our baseline experiments. Below we provide key observations, and also compare to the findings of Sikasote et al. [6], who evaluated gender bias for Bemba and Nyanja ASR on test sets without overlapping sentences between the male and female speech.

A. Bias Analysis for Bemba

We observe performance disparities across all speaker attribute categories for all fine-tuned models. All models perform better in recognizing speech for native than non-native speakers, suggesting potential bias against non-native speakers in the models. All models perform better in recognizing male than female speech with relatively low CER-gap compared to other attributes. In contrast, Sikasote et al. [6] found inconsistent bias patterns between two evaluation sets both without sentence overlap between the subgroups.

B. Bias Analysis for Nyanja

We observe significantly higher CER and CER-gap values in Nyanja than in Bemba. We attribute this to the possible impact of background noise. Our speech analysis revealed that approximately 1.5 hours of the total Nyanja data is non-speech compared to Bemba, with only 45 minutes. Similarly to Bemba for the *Speaker Type* attribute, all models perform better in recognizing speech for native speakers than non-native speakers, suggesting potential bias against non-native speakers in the models. For gender, all fine-tuned models exhibit bias against male speakers, consistent with the pattern observed in the study by Sikasote et al. [6].

VI. CONCLUSION

This paper introduces AfriBED, a bias evaluation dataset created to examine gender and speaker-type bias in ASR systems for two low resource African languages: Bemba and Nyanja. The dataset's key innovation is that the subgroups of each speaker attribute have the same speech content, making the subgroups comparable for bias evaluation. The results of our baseline experiments demonstrate the usefulness of our datasets to examine gender and speaker-type attribute bias in ASR models for Bemba and Nyanja. In the future, we plan to expand our dataset to more languages, and to add additional speaker attributes.

Acknowledgment

This project would not have been successful without the voluntary participation of students and the incredible support of academic staff in the Department of Computing and Informatics at the University of Zambia. We express our profound gratitude to Mofya Phiri, Monica M. Kabemba, and Alinani Simukanga for the useful conversations and support during the course of the data collection process.

This work was financially supported by the Hasso Plattner Institute (HPI) for Digital Engineering, through the HPI Research School at the University of Cape Town. This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number 151601) and Telkom Centre of Excellence. Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing (HPC) team: https://ucthpc.uct.ac.za/

AI-Generated Content Disclosure

We used ChatGPT to generate LaTeX code to format and draw tables. The writing, technical content, experiments, and conclusions were produced and verified by the authors.

References

[1] S. H. Imam, B. Sani, D. K. Gete, B. Y. Ahmed, I. S. Ahmad, I. Abdulmumin, S. M. Yimam, M. Y. Bello, and S. H. Muhammad, "Automatic speech recognition for african low-resource languages: Challenges and future directions," in *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pp. 89–94, 2025.

[2] C. Sikasote and A. Anastasopoulos, "BembaSpeech: A speech recognition corpus for the Bemba language," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (Marseille, France), pp. 7277–7283, European Language Resources Association, June 2022.

[3] C. Sikasote, K. Siaminwe, S. Mwape, B. Zulu, M. Phiri, M. Phiri, D. Zulu, M. Nyirenda, and A. Anastasopoulos, "Zambezi Voice: A Multilingual Speech Corpus for Zambian Languages," in *Interspeech 2023*, pp. 3984–3988, 2023.

[4] C. Sikasote, E. Mukonde, M. M. I. Alam, and A. Anastasopoulos, "BIG-C: a multimodal multi-purpose dataset for Bemba," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 2062–2078, Association for Computational Linguistics, July 2023.

[5] E. N. Um, F. Tyers, E.-C. E. N. Tjomb, F. L. Dibengue, B.-M. B. Manguele, B. A. Djoulde, M. Nyambe, B. M. A. Eloundou, J. S. N. Kamagoua, J. M. Avom, *et al.*, "Speech technologies datasets for african under-served languages," in *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 82–90, 2025.

[6] C. Sikasote, H. Suleman, and J. Buys, "Investigating the impact of multilingual pre-trained speech models on gender bias in asr for low-resource african languages," in *Artificial Intelligence Research* (A. Gerber and A. W. Pillay, eds.), (Cham), pp. 144–158, Springer Nature Switzerland, 2026.

[7] K. Siminyu, K. Mohamed Amran, A. Ndegwa Karatu, M. Resani, M. Makobo Junior, R. Ryakitimbo, and B. Mwasaru, "Corpus development of kiswahili speech recognition test and evaluation sets, preemptively mitigating demographic bias through collaboration with linguists," in *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages* (S. Moeller, A. Anastasopoulos, A. Arppe, A. Chaudhary, A. Harrigan, J. Holden, J. Lachler, A. Palmer, S. Rijhwani, and L. Schwartz, eds.), (Dublin, Ireland), pp. 13–19, Association for Computational Linguistics, May 2022.

[8] J. Nakatumba-Nabende, S. Kagumire, C. Kantono, and P. Nabende, "A systematic literature review on bias evaluation and mitigation in automatic speech recognition models for low-resource african languages," *ACM Comput. Surv.*, vol. 58, Oct. 2025.

[9] T. Kendall and C. Farrington, "The corpus of regional african american language (coraal)," 2023. Available: https://doi.org/10.7264/1ad5-6t35. Accessed: 2026-07-02.

[10] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Marseille, France), pp. 6462–6468, European Language Resources Association, May 2020.

[11] I.-E. Veliche, Z. Huang, V. Ayyat Kochaniyan, F. Peng, O. Kalinli, and M. L. Seltzer, "Towards measuring fairness in speech recognition: FairSpeech dataset," in *Interspeech 2024*, pp. 1385–1389, 2024.

[12] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, and Y. Saraf, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6162–6166, 2022.

[13] M. Jahan, Y. Sun, P. Mazumdar, Z. Fagyal, T. Thebaud, J. Villalba, M. Hasegawa-Johnson, N. Dehak, and L. M. Velazquez, "FaiST: A Benchmark Dataset for Fairness in Speech Technology," in *Interspeech 2025*, pp. 1343–1347, 2025.

[14] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Marseille, France), pp. 4218–4222, European Language Resources Association, May 2020.

[15] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, (Cambridge, Massachusetts, USA), pp. 97–100, Association for Computational Linguistics, Apr. 30 - May 1 2004.

[16] Zambia Statistics Agency, "2022 census of population and housing: National analytical report." National Analytical Report, July 2025. Available: https://www.zamstats.gov.zm/wp-content/uploads/2025/08/2022-Census-National-Analytical-Report.pdf. Accessed: Jul. 8, 2026.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5206–5210, IEEE, 2015.

[18] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2023.

[19] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," 2023.

[20] A. Caubrière and E. Gauthier, "Africa-centric self-supervised pretraining for multilingual speech representation in a sub-saharan context," in *5th Workshop on African Natural Language Processing*, 2024.

[21] M. Kurimo, M. Creutz, M. Varjokallio, E. Arsoy, and M. Saraclar, "Unsupervised segmentation of words into morphemes - morpho challenge 2005 application to automatic speech recognition," in *Interspeech 2006*, pp. paper 1512–Tue2A2O.1, 2006.

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