Towards inclusive automatic speech recognition

Feng, Siyuan; Halpern, Bence Mark; Kudina, Olya; Scharenborg, Odette

DOI
10.1016/j.csl.2023.101567

Publication date
2023

Document Version
Final published version

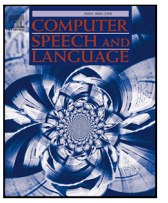
Published in
Computer Speech and Language

Computer Speech and Language 84 (2024) 101567

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl



Towards inclusive automatic speech recognition

Siyuan Feng [a], Bence Mark Halpern [a,b,c], Olya Kudina [d], Odette Scharenborg [a,*]

[a] Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
[b] Netherlands Cancer Institute, Amsterdam, The Netherlands
[c] ACLC, University of Amsterdam, Amsterdam, The Netherlands
[d] Ethics and Philosophy of Technology Section VTI Department, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Keywords:
Inclusive automatic speech recognition
Bias
Gender
Age
Accent

ABSTRACT

Practice and recent evidence show that state-of-the-art (SotA) automatic speech recognition (ASR) systems do not perform equally well for all speaker groups. Many factors can cause this bias against different speaker groups. This paper, for the first time, systematically quantifies and finds speech recognition bias against gender, age, regional accents and non-native accents, and investigates the origin of this bias by investigating bias cross-lingually (i.e., Dutch and Mandarin) and for two different SotA ASR architectures (a hybrid DNN-HMM and an attention based end-to-end (E2E) model) through a phoneme error analysis. The results show that only a fraction of the bias can be explained by pronunciation differences between speaker groups, and that in order to mitigate bias, language- and architecture specific solutions need to be found.

1. Introduction

Automatic speech recognition (ASR) is increasingly used, in, e.g., emergency response centers, domestic voice assistants, and search engines. Because of the paramount relevance spoken language plays in our lives, it is critical that ASR systems are able to deal with the variability in the way people speak (e.g., due to speaker differences, demographics, and differently abled speakers).

State-of-the-art (SotA) ASR systems are based on deep neural networks (DNNs). DNNs are often considered to be a harbor of objectivity because they follow a clear path against the set parameters applied to the provided dataset. Although studies on bias in ASR are only nascent, practice and recent evidence are already troubling, suggesting that the SotA ASR systems do not recognize the speech of everyone equally well. This evidence ranges from anecdotal (e.g., the smart speaker of author O.S. does not recognize the speech of her 9-year-old daughter) to research- and policy-oriented. For instance, ASR systems have been shown to struggle with speech variance due to gender, age, speech impairment, race, and accents. Studies across languages have repeatedly found recognition bias between genders, predominantly favoring female speakers (Arabic Abu Shariah and Sawalha, 2013, English Koenecke et al., 2020; Adda-Decker and Lamel, 2005; Goldwater et al., 2010, and French Adda-Decker and Lamel, 2005), while male speech was best recognized in other studies (French Garnerin et al., 2019, English Tatman, 2017), although a follow-up study to the latter study found no difference between genders (Tatman and Kasten, 2017) nor was a difference found in Garnerin et al. (2019). It should be noted that these studies do not include transgender and non-binary speakers.

Speakers younger than 30 years of age are better recognized than those older than 30 years (Abu Shariah and Sawalha, 2013), while the recognition of child speech is more challenging than that for adult speech, due to children's shorter vocal tracts, slower and more variable speaking rate and inaccurate articulation (Qian et al., 2017). A speech impairment, e.g., due to dysarthria (Moro-Velázquez et al., 2019), stroke survival, oral cancer (Halpern et al., 2020) or cleft lip and palate (Schuster et al., 2006), is known to

* Corresponding author.
E-mail addresses: fengsym.ee@gmail.com (S. Feng), b.halpern@nki.nl (B.M. Halpern), o.kudina@tudelft.nl (O. Kudina), o.e.scharenborg@tudelft.nl (O. Scharenborg).

https://doi.org/10.1016/j.csl.2023.101567
Received 10 March 2022; Received in revised form 1 September 2023; Accepted 7 September 2023
Available online 20 September 2023

cause many problems for standard ASR systems. Recent studies further demonstrate how voice assistants perpetuate a racial divide by misrecognizing the speech of African American speakers more often than of white speakers (Koenecke et al., 2020; Tatman and Kasten, 2017). Finally, ASR systems are typically trained on speech from native speakers of a "standard" variant of that language, inadvertently discriminating not only the speech of non-native speakers (Wu et al., 2020; Palanica et al., 2019) but also that of speakers of regional or sociolinguistic variants of the language (English Koenecke et al., 2020; Tatman, 2017; Tatman and Kasten, 2017, Arabic Abu Shariah and Sawalha, 2013).

There are many factors that can cause this bias, and different locations in the ASR system where these factors manifest themselves. Such bias-inducing factors, for instance, include, (1) under-representation of the speaker group in the training data (i.e., the composition of the training data). This leads to acoustic models (AMs) that will not be able to capture the pronunciations of that speaker group well. (2) Within-group variability: Even if the ASR is trained only on speech of the underrepresented group, recognition performance is often found to be worse due to the large variability both in the pronunciation and in language use within the speaker group (e.g., Koenecke et al., 2020; Tatman and Kasten, 2017; Qian et al., 2017). (3) The transcriptions can be biased. Anecdotal evidence from author B.M.H. on the Jasmin-CGN corpus (Cucchiarini et al., 2006) suggests that production errors of children are corrected ("normalized" towards what should have been said) in a more lenient way than those of non-native adult speakers (transcriptions tend to be more verbatim, including restarts). Moreover, transcriptions might be less accurate because the annotators have less experience with the type of speech. (4) Across-group variability: A speaker group that has a dialect that deviates significantly from that of the other speaker groups in the training data is usually recognized worse (Winata et al., 2020; Alsharhan and Ramsay, 2020). (5) Not all speaker groups might have access to equally high-quality recording equipment. (6) Possibly, bias can be due to the specific ASR architectures, of which there are two main categories in current ASR: end-to-end (E2E) and hybrid Deep Neural Network (DNN)-hidden Markov model (DNN-HMM), and algorithms used in ASR system development. (7) Bias also creeps in far before the datasets are collected and deployed, e.g., when framing the problem, preparing the data and collecting it (e.g., Caliskan et al., 2017). Most of these factors will have their impact on the acoustic model (AM), e.g., leading to a mismatch with the trained AM. However, deviant language use will also have an effect on the language model.

Our goal is to create inclusive ASR, i.e., ASR for everyone, irrespective of how one speaks or the language one speaks (Scharenborg, 2021). As first and crucial steps towards this larger goal, here, we quantify bias in state-of-the-art ASR systems, and investigate the origin of this bias by investigating bias (1) against different speaker groups or dimensions; (2) in two different speaking styles in order to answer the question whether the size of the bias is influenced by the speaking style of the person; (3) cross-lingually in two vastly different languages (non-tonal Dutch and tonal Mandarin) in order to answer the question whether bias is language dependent; and (4) for different SotA ASR architectures (a hybrid DNN-HMM and an attention based E2E model) in order to answer the question whether bias is dependent on the ASR architecture. The results will allow us to work towards proactive bias-mitigation in ASR systems.

Prior work in the literature typically focused on one to three speaker groups or dimensions, here we will investigate possible bias against gender, age, regional accents, and non-native accents. In our search for the origin of the bias, we carry out an analysis of which sounds are particularly prone to misrecognition under the assumption that different speaker groups exhibit different global patterns of pronunciation variation and that phoneme error patterns might be indicative of particular problems that lead to bias, e.g., misrecognized vowels or voiced final obstruents misrecognized as their unvoiced counterparts might be indicative of regional or non-native speech.

2. Speech database selection and design

In order to be able to quantify bias for different speaker groups, we are crucially dependent on the meta-data available in the speech databases. We therefore carefully selected and curated our speech databases. For Mandarin, we only found databases that allowed us to investigate bias against gender and regional accents.

2.1. Dutch corpora

2.1.1. Dutch spoken corpus (CGN)

The CGN corpus (Oostdijk, 2000) is used to train the Dutch SotA ASR systems. CGN contains Dutch recordings spoken by 1185 female and 1678 male speakers (age range 18–65 years old) from all over the Netherlands (NL) and Flanders (FL, in Belgium (BE)). It contains 14 different speaking styles. In this study, only CGN data from NL is used for training, while both Dutch from NL and FL are used for testing. We used the standard training and test sets (Leeuwen et al., 2009). Two test sets were used, one for each speaking style: broadcast news (BN) and conversational telephone speech (CTS). All recordings were first pre-processed by cutting the speech signals into smaller chunks and removing the silence chunks. Table 1 presents detailed information about the CGN training and test sets after the pre-processing steps.

2.1.2. Jasmin-CGN corpus

The Jasmin-CGN corpus (Cucchiarini et al., 2006), which is an extension of the CGN corpus, is used to evaluate the Dutch ASR systems' bias against gender, age, regional and non-native accent. The recording conditions of CGN and Jasmin-CGN are different, which might lead to an ASR performance deterioration on Jasmin-CGN compared to CGN. However, in our bias investigations we only compare speaker groups within the Jasmin-CGN dataset, avoiding this potential problem. We use the speech from the following groups: (1) DC: Dutch children; age 7–11 years; (2) DT: Dutch teenagers; age 12–16; (3) DOA: Dutch older adults; age 65+; (4) NNC:

Table 1
Hours of speech data and numbers of speakers in the CGN training and test sets for Dutch.

	Training		BN test		CTS test	
	#Hrs	#Spks	#Hrs	#Spks	#Hrs	#Spks
All	423	2863	0.4	4	1.8	25
Female	193	1185	0.2	1	0.8	12
Male	230	1678	0.2	3	1.0	13

Table 2
Number of native female and male speakers of Dutch per age group per NL region (D-) and for FL (F-) in the Jasmin-CGN corpus.

Region	W	T	N	S	FL
DC/FC	0, 0	15, 14	11, 11	9, 11	23, 19
DT/FT	9, 11	2, 2	10, 10	10, 9	22, 21
DOA/FOA	13, 5	9, 8	13, 4	10, 6	21, 16

non-native Dutch speaking children; age 7–16 years (28 female and 25 male speakers); (5) NNA: non-native adults; age 18–60 (28 female and 17 males speakers); with a wide range of native languages. The adults have different levels of proficiency of Dutch according to the Common European Framework (CEF; A1 the lowest): A1 (4 females, 6 males), A2 (18 females, 7 males), B1 (6 females, 3 males), B2 (1 male). The speakers come from four different regions in NL: W: West, T: Transitional, N: North, S: South. Moreover, we tested the ASR trained on NL Dutch on the speech of (1) FC: Flemish children; age 7–11; (2) FT: Flemish teenagers; age 12–16; (3) FOA: Flemish older adults; age 65+.

Table 2 shows the number of speakers broken down by gender (female, male[1]) for each age group and each (NL or FL) region, excluding the non-native speakers for which this information was not available. The Jasmin-CGN corpus consists of read speech and human–machine interaction (HMI) speech, both of which are used in the experiments. The number of hours for each region and age group ranges from 0.2 h (DT from region T) to 2.0 h (female FT) and from 0.2 h (several speaker groups from region S) to 0.9 h (FOA) for HMI speech. The number of hours of speech of the non-native speakers ranged from 0.2 h (B2 male speaker) to 0.8 h (A1 and A2 male speakers for read speech and from 0.1 h (A1 female speakers) to 1.1 h (A2 female speakers) for HMI speech.

2.2. Mandarin corpus

The MagicData Read Speech Corpus (Magic Data Technology Co., Ltd., 2019) is an open-source Mandarin speech corpus consisting of 755 h of Mandarin recordings spoken by adult speakers (age range 18–55 years old) from seven regions from all over mainland China: Northern Guan (NG), Southern Guan (SG), Gan (GA), Min (MI), Wu (WU), Xiang (XI) and Yue (YU).[2] In the seven regions, speakers from NG and SG use a variety of Mandarin as their local languages, while local languages in GA, MI, WU, XI and YU are non-Mandarin Sinitic languages. The supplementary information A provides the mapping from a Chinese province to its accent region.

We followed the standard training, development and test data partitioning in MagicData, but with two necessary modifications: (1) The original test set does not contain 10 speakers for all accent regions. In order to avoid the results being dependent on the characteristics of individual speakers, we empirically set the minimum number of test speakers in every Chinese accent region at 10. To that end, speakers from the original training set were randomly selected and moved to the test set. (2) The original test set did not contain any female speech for NG and SG. In order to balance gender, the female speakers from the NG and SG regions in the original development set were moved to the test set. There is no speaker overlap between the training and test sets. Table 3 shows the number of hours of speech and test speakers broken down by gender and by Chinese accent regions in our test data.

2.3. Experiments and evaluation

In our experiments on Dutch, the potential bias due to gender, age, regional and non-native accents is quantified for read speech and HMI speech separately. For Mandarin, the bias against gender and regional accents is quantified for read speech only.

We quantify the bias of ASR systems for Dutch on the Jasmin-CGN corpus and for Mandarin on the MagicData corpus. We define bias as the difference in WER between the different speaker groups within each of the investigated dimensions, and it is computed by subtracting the lowest (=best) WER from the WER of each speaker group in the dimension. Moreover, for Dutch, for all dimensions, we split by age group. One-way analyses of variance (ANOVA) were carried out comparing the WER of each speaker group within each of the dimensions to investigate the significance of the bias, with per-speaker WER as the dependent variable, and each of the dimensions as the independent variable. We report the p-value and F-statistic.

[1] Please note that the meta-data only provides information regarding these two genders.

[2] Hakka is also a major Sinitic language; it is mainly spoken in the provinces of Guangdong, Guangxi and Fujian. However, Hakka is not included in this paper, as region information for speakers in MagicData is only available at the province level, making Hakka indistinguishable from Yue (Guangdong, Guangxi) and Min (Fujian).

Table 3
Hours of speech data and number of (male/female) speakers in the MagicData training and test sets. The test data is also broken down by accent region.

Set		All		Female		Male	
		#Hrs	#Spks	#Hrs	#Spks	#Hrs	#Spks
Training		680.1	968	366.9	516	313.2	452
Test	All	52.1	78	26.5	37	25.6	41
	NG	11.1	16	8.4	11	2.7	5
	SG	8.7	12	2.4	3	6.3	9
	GA	6.5	10	2.8	4	3.7	6
	MI	7.0	10	3.8	5	3.2	5
	WU	5.6	10	2.3	4	3.2	6
	XI	6.5	10	2.9	4	3.6	6
	YU	6.7	10	4.0	6	2.8	4

In order to understand the source of the bias, we computed phoneme error rates (PER) for each individual phoneme to investigate whether certain phonemes are prone to misrecognitions. The PER is calculated similarly to the calculation of the WER but using phoneme-level transcriptions (converted from word-level transcriptions using lexicons) of the reference and hypothesized word sequences.[3]

2.4. The state-of-the-art hybrid and E2E ASR systems

For the experiments, we used a SotA factorized TDNN (Povey et al., 2018) (TDNNF) implemented using Kaldi (Povey et al., 2011)[4] as our hybrid model and a conformer-based encoder–decoder model implemented using ESPnet as our E2E model, one for Dutch and one for Mandarin (Watanabe et al., 2018). For both architectures, the same training material and MFCC acoustic feature representations were used.

2.4.1. Hybrid DNN-HMM architecture

The TDNN-BLSTM model for both the Dutch and Mandarin ASR systems consisted of three TDNN layers of dimension 1024, and 3 pairs of forward–backward LSTM layers of cell dimension 1024 on top. The TDNNF model for the Dutch and Mandarin ASRs consisted of 12 TDNNF layers of dimension 1024. For the Mandarin TDNNF model, we also added 6 convolutional layers between the input layer and the first TDNNF layer, following the recommended layout.[5]

The language model (LM) in the hybrid ASR system is an RNNLM (Xu et al., 2018). It consists of 3 TDNN layers interleaved with 2 LSTM layers. The RNNLM is trained with 20 epochs. To apply the RNNLM, a tri-gram LM is used to generate N-best results. After that, the RNNLM rescores the N-best results to get the final recognition results. The RNNLM and the tri-gram LM are trained using the training data transcriptions in CGN for Dutch and MagicData for Mandarin.

2.4.2. End-to-end (E2E) architecture

The conformer E2E model parameters were mainly taken from Guo et al. (2021) and Karita et al. (2019): 12 encoder layers and 6 decoder layers, all with 2048 dimensions; the attention dimension is 512 and the number of attention heads is 8; the convolution subsampling layer in the encoder has 2-layer CNNs with 256 channels, stride with 2, and a kernel size of 3. The default kernel size (31) of the CONV module in the conformer structure was used for the Dutch ASR, while a CONV kernel size of 15 was used[6] for the Mandarin ASR. The conformer model was trained with 50 epochs using a joint connectionist temporal classification (CTC)-attention objective (Kim et al., 2017), in which the CTC and attention weights were set to 0.3 and 0.7, respectively. For the Dutch conformer model, subword units with a vocabulary size of 5000 were used as basic units. For the Mandarin conformer model, Chinese characters with a vocabulary size of 4481 were used as basic units.

An RNNLM was trained for each language, and used during E2E ASR decoding in a shallow fusion manner (Hori et al., 2017). The RNNLM consisted of 2 LSTM layers of dimension 1024, and was trained with the training data transcripts of CGN (Dutch) or MagicData (Mandarin) for 40 epochs.

[3] Source code of the analysis method can be found at: https://github.com/karkirowle/relative_phoneme_analysis.

[4] A preliminary experiment compared a TDNN-BLSTM (Feng and Lee, 2018) and a factorized TDNN (Povey et al., 2018) (TDNNF) model, both were implemented using Kaldi (Povey et al., 2011), used the same training material and the same MFCC acoustic features. Although the TDNN-BLSTM outperformed the TDNNF system on the in-domain CGN BN set, the TDNN-BLSTM performed worse than the TDNNF on the out-domain Jasmin-CGN corpus, our test corpus. For Mandarin, we also observed that the TDNNF model outperformed the TDNN-BLSTM model. Therefore, in our experiments, we used the TDNNF system. Details of the in-domain and out-domain WER results for Dutch are listed in Tables B.1 and B.2.

[5] `run_cnn_tdnn_1b.sh` in the Kaldi `multi_cn` recipe.

[6] This parameter is recommended in the ESPnet recipe of `aidatatang_200zh`.

Table 4
Bias sizes for the TDNNF hybrid and conformer E2E ASR systems for gender (4a), age (4b), and regional accents (4c) split by age group on the Jasmin-CGN read and HMI speech. The calculation method of the bias and the group with the best recognition performance is indicated in the respective subcaptions.

(a) Bias against gender. A positive bias means that female speech was recognized better than male speech.

	Read		HMI	
	Hybrid	E2E	Hybrid	E2E
DC	0.3	−0.5	−3.0*	−0.4
DT	2.5*	3.2**	1.5	3.7*
DOA	5.1	5.3*	5.1	6.5*
NNC	1.1	1.8	1.0	4.4
NNA	0.4	1.3	3.7	5.6

(b) Bias against age group. For the Dutch speakers, teenagers were recognized best; for the non-native speakers, child speech was recognized best.

	Read		HMI	
	Hybrid	E2E	Hybrid	E2E
DC	11.8***	12.2***	7.2***	7.2***
DT	0.0	0.0	0.0	0.0
DOA	4.6**	3.9*	7.8***	8.2***
NNC	0.0	0.0	0.0	0.0
NNA	1.6	1.9	1.2	1.6

(c) Bias against native regional accents. Bias size numbers are calculated by subtracting the lowest WER (first field within brackets) from the highest WER (second field within brackets) among five regions (including FL) within an age group. "Avg" indicates the average bias size over all age groups.

	Read		HMI	
	Hybrid	E2E	Hybrid	E2E
Regions: W, T, N, S, FL				
DC/FC	11.5*** (T,FL)	12.6*** (T,FL)	20.5*** (N,FL)	19.6*** (N,FL)
DT/FT	16.4*** (N,FL)	20.9*** (N,FL)	15.8* (T,FL)	19.9** (T,FL)
D/FOA	11.7** (N,S)	12.2** (N,S)	13.1* (N,S)	13.8*** (N,FL)
Avg	13.2	15.2	16.5	17.8

* $p < .05$.
** $p < .01$.
*** $p < .001$.

2.4.3. Dutch in-domain ASR performances for reference

The TDNNF and E2E models were first evaluated on the in-domain CGN data, specifically on the BN test set, which is closest in speaking style to the read speech in the Jasmin-CGN corpus, and the CTS test set, which is closest to the HMI speaking style in the Jasmin-CGN corpus. Comparing these numbers with the native speaker results on Jasmin-CGN provides an indication of the performance drop due to the differences in database.

On BN speech, the TDNNF system (6.3%) slightly outperformed the E2E system (6.6%), while the E2E system outperformed the TDNNF (21.6% vs. 23.9%) on CTS speech. Details of the in-domain WER results are listed in Table B.1.

3. Quantifying bias

3.1. Bias in state-of-the-art ASRs for Dutch

3.1.1. Bias against gender

Overall, female speech was recognized similarly or better than male speech (see Table B.3 in the supporting materials for the WER breakdown for gender, age and non-nativeness). Table 4a lists the size of the bias against male speakers compared to female speakers split for native and non-native speakers and for the different age groups, for the hybrid and E2E systems and the read speech and HMI speech test sets.

The native speaker groups. For the hybrid models, we only observe bias for two cases: for read speech, male teenagers are significantly worse recognized than female teenagers ($F(1,61) = 5.543$, $p = .022$), while for HMI speech, female children are significantly worse recognized than male children ($F(1,69) = 4.316$, $p = .041$). The E2E ASR is more prone to bias. We observed a statistically significant bias against male speakers for teenagers (read speech: $F(1,61) = 7.953$, $p = .006$; HMI speech: ($F(1,61) = 4.036$, $p = .049$)) and older adults (read speech: $F(1,66) = 4.122$, $p = .046$; HMI speech: ($F(1,66) = 6.350$, $p = .014$)) in both speech styles. No biases were observed for the **non-native speaker groups**.

Both architectures thus exhibited a bias against male speakers, however this bias was much less for the hybrid model compared to the E2E model. No gender bias was observed for the non-native listeners. This finding could however be due to the relatively high WERs for the non-native speaker groups. These results add to a growing set of findings that male and female speech are not

recognized equally well (Koenecke et al., 2020; Tatman, 2017; Abu Shariah and Sawalha, 2013; Adda-Decker and Lamel, 2005; Goldwater et al., 2010).

3.1.2. Bias against age

Overall, for the native speakers, speech from teenagers was recognized best, followed by that of older adults, while child speech was recognized worst (see Table B.3 for the WER breakdown for gender, age and non-nativeness). For the non-native speakers, child speech was recognized better than that of adult speakers. Table 4b lists the size of the bias against native children's and older adults' speech (top rows) and against non-native adults' speech (bottom row), for read and HMI speech, and the hybrid and E2E models, separately.

The native speakers groups. We observe substantial age bias: speech from teenagers was found to be significantly better recognized than that of children for both models and both speaking styles (hybrid ASR on read speech: ($F(1132) = 87.158$, $p < .001$); on HMI speech: ($F(1132) = 19.425$, $p < .001$); E2E ASR on read speech: ($F(1132) = 88.815$, $p < .001$); on HMI speech: ($F(1132) = 25.691$, $p < .001$)) and significantly better recognized than speech from older adults for both models and both speaking styles (hybrid ASR on read speech: ($F(1129) = 7.573$, $p = .007$); on HMI speech: ($F(1129) = 15.804$, $p < .001$); E2E ASR on read speech: ($F(1129) = 6.533$, $p = .012$); on HMI speech: ($F(1129) = 18.935$, $p < .001$)).

The age bias size, though, is different for the two speech styles: the bias against children's speech is smaller for HMI than for read speech, while the bias against older adults' speech is larger for HMI than for read speech. Informal listening to a few of the child speakers' recordings suggests that the smaller bias for HMI speech is due to a higher WER on read speech, and could be due to high volume and disfluency/hesitations in the children's read speech. No bias was observed for the **non-native adult speakers** compared to the speech of that of non-native children.

In conclusion, both architectures exhibited a (large) age bias against children's and older adults' speech for native speakers of Dutch, while no age bias was observed for the non-native speakers. The size of the bias seems to be similar for the two architectures. The problems of the ASR with recognizing children's speech can be explained by the large difference in children's speech and adults' speech (Qian et al., 2017) which leads to a large mismatch of the children's speech with the AM. The worse recognition of the older adults' speech, especially those over 75 y/o, is likely due to a less well articulation.

3.1.3. Bias against non-native accents

The speech of native speakers was recognized better than that of non-native speakers of Dutch (Table B.3). The bias against non-native accents is significant for both speaking styles and both architectures (hybrid ASR on read speech: Size = 23.1; $F(1298) = 282.851$, $p < .001$; on HMI speech: Size = 13.9; $F(1298) = 126.716$, $p < .001$; E2E ASR on read speech: Size = 24.5; $F(1298) = 344.457$, $p < .001$; on HMI speech: Size = 16.7; $F(1298) = 197.807$, $p < .001$). The hybrid system seemed to exhibit a smaller bias against the non-native speakers than the E2E architecture.

These results are in line with the qualitative findings reported in Wu et al. (2020) and Palanica et al. (2019). Non-native speakers typically have an accent, meaning that the match with the AM is worse than that of native speakers. For the non-native speakers, on average, the WER results by both models showed a decrease when CEF level increases (see Table B.4; except for the one B1 speaker for read speech). This is in line with the intuition that non-native speakers with a higher CEF level tend to speak Dutch better than those with a lower level.

3.1.4. Bias against regional accents

Overall, there is a large variety in the recognition performance of the speech from the different accent regions in the Netherlands and Flanders, with speech from Flanders recognized worst (see Table C.1 for the WER breakdown per accent region). Table 4c lists the size of the bias against the five Dutch-speaking regions including Flanders, for every speaker age group, speech style, and ASR architecture, separately. Information in the brackets indicates which regions got the lowest and highest CER in every age group. All biases were shown to be significant ($p < .029$). For all age group, a bias in regional accents was observed. This finding is due to the fact that FL speakers were much worse recognized than any NL region's speakers regardless of age (see Table C.1), which in turn is likely due to the lack of the use of FL training speech data.

In conclusion, both architectures showed clear biases against regional accents, particularly FL. This bias was similar for the hybrid system compared to the E2E system.

3.1.5. Summary of bias in state-of-the-art ASRs for Dutch

The results showed that the Dutch ASR have (1) a gender bias, with a bias against male speech; (2) an age bias for native speakers of Dutch, with the largest bias against speech of native children, followed by speech of native older adults; (3) a bias against non-native speech, with an absolute WER degradation of around 24% in recognizing non-native speakers' read speech and 15.0% in HMI speech; and (4) a bias against regional accents with the strongest biases against Flemish and speech from the south of the Netherlands. Comparing the biases exhibited by the two ASR architectures showed similar or smaller biases for the hybrid ASR system compared to the E2E system.

3.2. Bias in state-of-the-art ASRs for Mandarin

Unlike in the Dutch ASR experiments, there is no domain mismatch between training and test data in the Mandarin ASR experiments. The TDNNF hybrid system and the conformer E2E system achieved overall CER results of 3.3% and 2.9% respectively, on the MagicData test set.

Table 5
Bias sizes for the TDNFF hybrid and conformer E2E ASR for gender (5a) and regional accents (5b) on MagicData.

(a) Bias against gender, split by region. Female speech was recognized best.

Set	Hybrid		E2E	
	Size	*P*-value	Size	*P*-value
All	0.4	.419	0.4	.315
NG	−0.2	.593	−0.5	.377
SG	−0.1	.744	0.2	.634
[a]GA	−0.1	.761	0.2	.621
[a]MI	2.6	.054	2.2	.104
[a]WU	1.1	.197	1.3	.087
[a]XI	0.3	.827	0.0	.742
[a]YU	0.5	.744	1.1	.366

(b) Bias against regional accents. Bias sizes are calculated by subtracting the CER of GA from the CER of itself.

	Hybrid		E2E	
	Size	*P*-value	Size	*P*-value
NG	0.8	.176	0.7	.266
SG	0.4	.112	0.3	.467
[a]MI	2.7	.001	2.3	.005
[a]WU	0.2	.592	0.1	.919
[a]XI	1.0	.012	0.7	.081
[a]YU	1.1	.085	1.0	.098

[a] The region uses non-Mandarin Sinitic languages as local languages.

3.2.1. Bias against gender

Overall, female speech was recognized slightly better than male speech (see Table C.2). Table 5a lists the size of the bias against male Mandarin speakers for each of the regions separately. No significant differences between the CER for the female and male speakers, thus no bias, was observed for both the hybrid and the E2E models.

3.2.2. Bias against regional accents

Overall, there is some variety in the recognition performance of the speech from the different accent regions, with speech from the Gan (GA) region being recognized best and that from the Min (MI) region recognized worst (see Table C.2 for all WERs).

Table 5b lists the size of the bias against the various regions compared to the best-recognized region GA, separately for both genders. For the hybrid system, the largest bias occurred against speakers from Min (MI) and Xiang (XI) ($F(1,18) = 14.165$, $p = .001$ and $F(1,18) = 7.757$, $p = .012$) respectively). For the E2E system, only a bias against MI speech was observed ($F(1,18) = 9.991$, $p = .005$).

In conclusion, both architectures showed a clear bias against MI (the worst recognized) speakers. Comparing the two ASR architectures shows that the E2E system was slightly less biased against regional (heavy) accents than the hybrid system, which also showed a bias against XI. Our finding regarding MI is in line with results reported in previous studies using a different database (Yi et al., 2018; Zheng et al., 2016).

3.2.3. Summary of bias in state-of-the-art ASRs for Mandarin

In summary, the results showed that our SotA Mandarin ASRs showed no bias against gender. Regarding regional accents, our two ASR systems were both biased against MI speakers, and the hybrid system was also biased against XI speakers. The E2E ASR system was less biased against regional (heavy) accents than the hybrid system.

4. Finding the origin of bias

4.1. Bias across speaker groups, architectures, speaking styles, and languages

The experiments showed that not only bias occurs in state-of-the-art ASR systems but also that it has many, different sources. First, bias occurs due to different aspects of the speaker as shown by the observed bias against the different genders, age groups, native vs. non-native speaker groups, and dialect speaker groups. We thus confirm and extend findings in the literature.

Second, bias and bias size are dependent on the architecture of the ASR system. For instance, we found a larger bias for the E2E models against male speakers for Dutch teenagers and older adults in both read speech and HMI speech, against non-native accents, against Flemish, and observed more bias against more strongly accented Mandarin.

Third, bias seems to be language-dependent. Although we can only compare bias against gender and region across Dutch and Mandarin, we do observe differences between the languages: while we found a bias against male speakers for Dutch for certain age groups, no gender bias was observed for the Mandarin speakers. Although distribution of speaker groups in the training data is often a cause of bias, here neither the biases nor the difference in bias between the languages can be explained by the distribution of these

Table 6
The five worst performing phonemes for each dimension that was found to have a significant bias for the Dutch ASRs for both architectures. Blue coloring indicates a discrepancy between the hybrid and E2E architectures. Red background coloring indicates that the phoneme is only worst recognized for a particular speaker group.

	Hybrid					E2E				
	1	2	3	4	5	1	2	3	4	5
Gender										
Female	ʒ	ʃ	œy	ʏ	y	ʒ	ʃ	ʏ	y	œy
Male	ʃ	ʒ	œy	ʏ	ɲ	ʒ	ʃ	ɲ	œy	ʏ
Age										
DC	ʏ	h	ə	j	y	ʏ	f	y	h	b
DT	ʃ	h	ʏ	ə	j	ʏ	ʃ	h	œy	ɔ
DOA	h	ɔ	ə	ʏ	f	ʒ	h	x	ɔ	ʃ
Native and non-native accents										
AvgD	ʒ	ʃ	ʏ	h	ə	ʒ	ʃ	ʏ	h	f
AvgNN	œy	y	ʒ	ʏ	ɲ	ʒ	œy	y	ʏ	h
Regional accents										
W	ʃ	h	ʏ	ə	j	h	ʏ	ʒ	ʃ	ə
T	ʃ	ʒ	ʏ	h	ə	ʒ	ʃ	ʏ	f	h
N	ʃ	ʒ	h	ʏ	ɔ	ʒ	ʃ	ʏ	ɔ	h
S	ʒ	ʏ	h	ə	j	ʏ	h	ʒ	ə	f
FL	ʃ	ʒ	œy	ɑu	ɛi	ʒ	ʃ	œy	ɑu	ʏ

two genders in the training material: For Dutch, the training data consisted of more male than female speech (male: 230 h; female: 193 h); the observed bias is thus against the speaker group with the most training data. For Mandarin, the training data consisted of substantially more female speech than male speech (male: 313.2 h; female: 366.9 h); this skewed distribution did not lead to a bias. A potential reason for the difference between the languages is that the gender bias for Dutch was analyzed on the out-domain Jasmin-CGN corpus, while for Mandarin this was analyzed on the in-domain MagicData corpus. Potentially, the matched train and test conditions for Mandarin underestimate the CER difference for the female and male speaker groups, or vice-versa, potentially, a mismatch in training and test conditions impacts male speech more than female speech.

Fourth, bias was observed for both speaking styles, but seems to occur slightly more often for more spontaneous speech. Potentially HMI speech, which is less well prepared than read speech allows for more speaker-dependent articulations and differences in word usage, which cause an increase in recognition problems for the ASR systems.

4.2. Phoneme analysis

We focus our phoneme analysis to increase our understanding of the origins of bias on read speech as bias seems to occur less often for read speech, so any results might also transfer across speaking styles. We compare across the two ASR architectures. We first identify the worst recognized phonemes for those dimensions that showed a significant bias for each language individually, and then compare the phoneme error patterns found for Dutch and Mandarin in order to find common patterns.

4.2.1. Dutch

Table 6 shows the breakdown of worst performing phonemes for the different dimensions (gender, age, non-nativeness, and native regional accents) for the two architectures separately. Blue coloring indicates a difference between the architectures; a red background coloring indicates a difference between speaker groups, for the specific phoneme.

The first thing to notice is the high similarity in the phonemes that are most difficult to recognize for each speaker group within each dimension: there are relatively few phonemes that are hardest to recognize that are not shared with the other speaker groups (not many phonemes with a red background). Also the relatively low number of blue colorings indicate that the architectures generally found the same phonemes hard to recognize. This is especially the case for gender, where four of the five worst recognized phonemes are shared. The bias against male speech can thus not be explained by a difference in pronunciation of specific phonemes which then would lead to specific phonemes being harder to recognize for male speech.

We observe a few more differences between the different age groups and between the two architectures. The hybrid system particularly seemed to have a problem recognizing the /ə/, while the phoneme pattern of most difficult to recognize phonemes differs somewhat between the age groups. The observed bias against the age groups can thus partially be explained by differences in pronunciation of specific sounds.

For the non-native speakers, we find that the observed phonemes are those, which are known to be challenging to acquire for second language speakers, such as /œy/ and /y/, therefore pronunciation differences between native and non-native listeners seem to be a factor in the origin of the bias. This conclusion is also corroborated by increasing CEFs levels showing decreasing WER in the hybrid ASR architecture.

Table 7
The five worst performing phonemes for the accent regions for the Mandarin ASRs.

	Hybrid					E2E				
	1	2	3	4	5	1	2	3	4	5
Native regional accent										
GA	z,	ɛ	au	a	ə	z,	au	a	ei	s
MI	z,	s	ə	au	ei	z,	s	ə	au	l
XI	z,	s	ə	au	a	z,	s	au	l	ə

Across the regional accents, we mostly see differences between Flanders (FL) and NL (W, T, N and S), where /œy/ and /ɑu/ are among the most problematic phonemes for both architectures. The biases observed for the FL speakers thus likely have a pronunciation-based origin. We further see that /ɔ/ is a difficult sound to recognize in N(orth) regional accent.

These results show that (general) pronunciation variation associated with specific speaker groups can lead to bias in automatic speech recognition. This bias is likely due to the pronunciation variation not being (fully) modeled by the acoustic models of our state-of-the-art ASR systems.

4.2.2. Mandarin

Table 7 lists the most problematic Mandarin phonemes for each regional accent. GA is the best recognized region whereas MI and XI are the worst recognized regions. It is clear that the bias against MI and XI cannot solely be explained by a difference in pronunciation of a range of phonemes: there is a large overlap in the phonemes that are worst recognized, except for the /s/. This latter finding can likely be explained by well-known variation patterns between the pronunciations of /ʂ/ and /s/, between /tʂ/ and /ts/, and between /tʂʰ/ and /tsʰ/ in Chinese regions using non-Mandarin Sinitic local languages (GA, MI, WU, XI and YU). We hypothesize that the high overall misrecognition of /s/ in MI and XI is caused by the /ʂ/ - /s/ ambiguity. Similar to the results found for Dutch, these results show that (general) pronunciation variation associated with specific speaker groups can lead to bias in automatic speech recognition.

4.2.3. General patterns

Comparing the phoneme patterns across the architectures shows that the hybrid and the E2E model show differences in which phonemes are hardest to recognize (blue phoneme symbols). In the case of the Dutch dataset, /j/ and /ə/ seem to be often poorly recognized by the hybrid system, while /f/ seems to be somewhat more difficult for the E2E system. For the Mandarin dataset, /l/ and /pʰ/ are more problematic for the E2E system. There thus are phoneme-specific differences the E2E and Hybrid systems.

Comparing the different speaker groups within the same architecture within the different dimensions, we see few differences for gender, age, and regional accents, for both Dutch and Mandarin Chinese. This shows that most bias cannot solely be explained by pronunciation differences.

4.3. Potential sources of bias

An often mentioned potential reason for bias in ASR systems is the composition of the training material. A skewed distribution of the speaker groups in the training data could not explain the gender bias (see Section 4.1). At a more fine-grained level, though, an analysis of the amount of training samples per phoneme and their error rates shows that sounds that are less frequent in the language and thus the training material (e.g., the /ʃ/, /ɲ/, /ʒ/ for Dutch) are typically less well recognized by the ASR systems. The composition of the training data thus plays a role in the recognition of specific phonemes, and in creating and ultimately removing bias.

The architectures tested in this work were trained on the same training material, but they still showed different patterns of bias. This suggests that the way the ASR architectures model the speech also plays a role in inducing bias, this could be at the level of the acoustic features, DNN architectures and/or training schemes: Evidence from a study on bias in speaker verification showed that the acoustic features used for processing the speech have shown to have limited capability to capture the speech characteristics of diverse speech (Hutiri and Ding, 2022; Li and Russell, 2001). Recent work from our group compared different training techniques including fine-tuning, multi-task training (Chen et al., 2015), and domain-adversarial training (DAT) (Sun et al., 2018). All showed performance improvements on standard speech but the results on non-native accented Dutch speech were less positive: fine-tuning led to a recognition performance worse than baseline for human–machine interaction speech (Zhang et al., 2022), while DAT showed little to no improvement in recognition performance of both native Dutch and non-native accented speech compared to standard training (Zhang et al., 2022). Moreover, the work presented here showed that different DNN architectures exhibit different biases, which warrants further investigation.

The analysis of the worst-recognized phonemes showed that only a fraction of the bias can be explained by pronunciation differences between speaker groups. This suggests that other sources of bias might exists. Potentially, differences between the speaker groups at the supra-segmental level induce bias, i.e., e.g., the fundamental frequency (F0), which is much higher in children's voices

than in adult's voices, speaking rate, which is typically slower in older adults than in younger adults, intonation, which differs substantially between Flemish and the Dutch Southern accent region and the other Dutch accent regions, and due to differences in word use.

5. General discussion and conclusion

Our goal is to uncover bias in state-of-the-art DNN-based ASR systems to work towards proactive bias-mitigation in ASR systems in order to create inclusive automatic speech recognition for everyone, irrespective of how they speak or the language they speak. In this paper, we have focused on bias that can be quantified. Our results show that bias occurs for different speaker groups, in different languages, and irrespective of the state-of-the-art ASR architecture that was used. The next steps are to further uncover the sources of this bias and finding ways to mitigate the bias.

It is important to note, that owing to the foundational nature of bias, it is impossible to remove bias that creeps into datasets (Kudina and de Boer, 2021). With this in mind, a priority in responsible ASR system development goes towards a proactive attitude. This concerns framing the problem, selecting the composition of the development team and the implementation process from a point of anticipating, proactively spotting, and developing both direct and indirect mitigation strategies for prejudice.

An indirect bias mitigation strategy deals with diverse team composition: the variety in age, regions, gender, etc. provides additional lenses of spotting potential bias in design. For example, according to the authors' (B.H.) experience diagnostic probing of the developed ASR systems often start with one-off audio recordings of the development team. These strategies together can help to ensure a more inclusive developmental environment for ASR.

There are potentially many different direct bias mitigation strategies that should be investigated in order to create truly inclusive ASR. The most obvious one concerns diversifying and aiming for a balanced representation of all types of speakers in the dataset (Koenecke et al., 2020; Caliskan et al., 2017). When a balanced representation of speakers is not possible or not sufficiently effective, data augmentation strategies could help in diversifying the dataset. For instance, recent work from our lab proposed cross-lingual voice conversion to create non-native accented Dutch which was successfully applied to improve recognition performance and reduce bias against non-native accented Dutch (Zhang et al., 2022). In line with the potential origins of bias as outlined in Section 4.3, other bias mitigation strategies should focus on improving the acoustic features, investigating different DNN training techniques, and investigating different DNN architectures. In short, there is no easy road to building inclusive ASR; rather, this research shows that many steps will need to be taken and that these steps should focus on all aspects of the ASR pipeline. Crucially, this research shows that we should not focus on blindly lowering the error rates on our test sets but that it is crucial to take into account the speaker groups and demographics that are inherently present in our test set and, more importantly, in society.

In conclusion, our research on only two languages already shows that there are big challenges to overcome in order to significantly reduce the bias in ASR systems, and these challenges are dependent on speaking style, language, and ASR architecture. Mitigation of bias should focus on improving the entire ASR pipeline, from training data composition, to acoustic features, DNN architectures and their training methods. Cross-language research will be important in highlighting where and how bias originates and how it can be reduced.

Declaration of competing interest

No interests to declare.

Data availability

The data and code is available through the URLs and references provided in the paper.

Acknowledgments

B.M.H. is funded through the EU's H2020 research and innovation program under MSC grant agreement No 766287.

Appendix A. Design of Mandarin-speaking regions

We grouped speakers in the MagicData Mandarin speech corpus into seven accent regions based on the province-level geographical information that was provided in the corpus for all speakers. The province name(s) contained in each region are listed below:

- **Northern Guan (NG)**: Beijing, Gansu, Hebei, Heilongjiang, Henan, Jilin, Liaoning, Neimenggu, Ningxia, Shandong, Shanxi, Tianjin, Xinjiang;
- **Southern Guan (SG)**: Anhui, Chongqing, Jiangsu, Guizhou, Hubei, Sichuan, Yunnan;
- **Gan (GA)**: Jiangxi;
- **Min (MI)**: Fujian;
- **Wu (WU)**: Shanghai, Zhejiang;
- **Xiang (XI)**: Hunan;
- **Yue (YU)**: Guangdong, Guangxi.

Table B.1
WERs of the TDNN-BLSTM and TDNNF hybrid ASRs and the conformer E2E ASR on the CGN standard broadcast news (BN) and conversational telephone speech (CTS) test sets. "F/M" indicates female/male. Numbers in bold indicate the best performance.

Arch.	Hybrid						E2E		
Model	TDNN-BLSTM			TDNNF			Conformer		
Set	Avg	F	M	Avg	F	M	Avg	F	M
BN	**5.6**	**5.5**	**5.6**	6.3	6.1	6.4	6.6	5.9	7.3
CTS	22.1	19.6	24.2	23.9	21.2	26.3	**21.6**	**18.8**	**24.0**

Table B.2
Average WERs over different age groups in Jasmin-CGN native and non-native speakers' read speech by the TDNN-BLSTM and TDNNF hybrid ASR systems.

	TDNN-BLSTM	TDNNF
Native (average)	30.0	19.6
Non-native (average)	59.5	42.7

Please note that in reality, more than one accents exist in some provinces. For instance, in Jiangsu, approximately 60% of the population uses SG, 30% of the population uses WU, and the rest uses NG (Wikipedia contributors, 2021). We decided to label Jiangsu as SG, as the majority of the speakers use SG in that region.

Appendix B. Word error rate details of the Dutch ASRs

B.1. In-domain results

Table B.1 lists the WER results on CGN (in-domain) test sets by the TDNN-BLSTM, TDNNF and conformer (E2E) systems.

B.2. Overall out-domain results

Table B.2 compared TDNN-BLSTM and TDNNF systems on out-domain (Jasmin-CGN) test data.

B.3. WER breakdown for gender, age, non-nativeness and regional accents

Table B.3 shows the WER per age group, for the female and male speech separately and averaged over both genders (column Avg), for read speech and HMI separately. The top rows list the results for the native Dutch speakers per age group; the bottom rows for the non-native speakers per age group. The WERs per gender, averaged over all age groups (row Avg), over the native (row AvgD) and non-native (row AvgN) Dutch speakers, respectively, are also shown.

B.3.1. WERs for per gender

Table B.3a and b show that, for both the hybrid and the E2E systems, in general, female speech was better recognized than male speech. Table B.3a and b also show that the average female-only and male-only read and HMI speech WERs achieved by the hybrid ASR system were all lower than the WERs by the E2E system.

B.3.2. WERs per age group

Table B.3a and b show that for both the hybrid and the E2E systems, among the native speakers, the Dutch teenager (DT) group achieved the best WER performances in read and HMI speech. Among the non-native speakers, the non-native children (NNC) group was slightly better recognized than the non-native adults (NNA) group. Comparing Table B.3a with Table B.3b shows that on both read and HMI speech, the hybrid system performed better than or similar to the E2E system in all the age groups.

B.3.3. WERs for native vs. non-native speakers

Table B.3a and b show that speech of native speakers was recognized much better than that of non-native speakers of Dutch, regardless of speech types and ASR systems. Comparing Table B.3a with Table B.3b also shows that on both read and HMI speech, the hybrid system performed better than or similar to the E2E system on native and non-native speech. Table B.4 provides a closer look at the WERs for the different Dutch proficiency levels (CEF) of all the non-native adult speakers (NNA), separated by gender. It shows a reduction in the overall WER (column Avg) with an increase in CEF level (except WERs on read speech by the E2E system). This is in line with the intuition that non-native speakers with a higher CEF level tend to speak Dutch better than those with a lower CEF level.

Table B.3
WERs of the TDNNF hybrid system (B.3a) and the conformer E2E system (B.3b) on the Jasmin-CGN read and HMI speech. "F/M" indicates female/male. "AvgD" indicates the average over all native Dutch speakers, "AvgN" over all non-native speakers and "Avg" indicates the average over all speakers.

(a) TDNNF hybrid system results

Group	Read			HMI		
	F	M	Avg	F	M	Avg
DC	25.6	25.9	25.8	31.5	28.5	30.0
DT	12.8	15.3	14.0	22.0	23.5	22.8
DOA	16.9	22.0	18.6	28.7	33.8	30.6
AvgD	18.3	21.1	19.6	28.4	30.8	29.4
NNC	41.5	42.6	42.0	42.0	43.0	42.5
NNA	43.4	43.8	43.6	42.2	45.9	43.7
AvgN	42.5	43.1	42.7	42.2	44.9	43.3
Avg	26.7	28.2	27.4	33.3	35.9	34.4

(b) Conformer E2E system results

Group	Read			HMI		
	F	M	Avg	F	M	Avg
DC	28.5	28.0	28.3	29.9	29.5	29.7
DT	14.5	17.7	16.1	20.6	24.3	22.5
DOA	18.3	23.6	20.0	28.2	34.7	30.7
AvgD	20.3	23.2	21.6	27.6	31.7	29.4
NNC	44.4	46.2	45.2	42.7	47.1	44.9
NNA	46.6	47.9	47.1	44.3	49.9	46.5
AvgN	45.5	46.9	46.1	43.9	49.0	46.1
Avg	29.1	30.8	29.8	33.4	38.0	35.3

Table B.4
WERs of the TDNNF hybrid system (Table B.4a) and the conformer E2E system (Table B.4b) on the Jasmin-CGN non-native (NNA) speaker group separated by CEF proficiency in Dutch levels (A1 is the lowest level). "F/M" indicates female/male. The one B2-level speaker was omitted from the NNA speaker group for this analysis.

(a) TDNNF hybrid system results

CEF	Read			HMI		
	F	M	Avg	F	M	Avg
A1	44.6	44.4	44.5	43.7	47.6	47.0
A2	44.9	38.7	43.3	44.4	41.4	43.5
B1	37.6	51.5	42.6	38.4	44.7	40.4

(b) Conformer E2E system results

CEF	Read			HMI		
	F	M	Avg	F	M	Avg
A1	48.3	46.9	47.5	45.7	50.8	49.6
A2	46.4	44.5	45.9	46.1	46.2	46.1
B1	46.1	54.8	49.1	41.2	50.3	43.5

B.3.4. WERs per native regional accents

Table C.1 shows the WERs for each of the regional accents of the four large regions W, T, N and S in the Netherlands and Flanders (FL) in Belgium per age group, by the hybrid system (C.1a and C.1b) and by the E2E system (C.1c and C.1d). The average WER results over female and male speakers are shown in the gray rows, and the results broken down by female and male are shown in the white rows.

Table C.1 shows that for both the hybrid and the E2E systems, speech spoken by people from Flanders (FL) achieved the worst performance in all age groups except for the older adults (DOA/FOA). Among the four regions in NL, for read speech, no region was consistently recognized worse than others; for HMI speech, region S in general was the worst recognized. Table C.1 also shows that on read speech, the hybrid system performed better than the E2E system in all the four regions in NL and the region FL; on HMI speech, no superiority of one ASR system over the other was observed in the four regions in NL, while the hybrid system was found better than the E2E system in FL.

Table C.1
WERs of the TDNNF hybrid system (C.1a and C.1b) and the conformer E2E system (C.1 and C.1b) on the Jasmin-CGN read and HMI speech of the four Dutch (NL) regions and Flanders in Belgium (BE) per age group. The average WERs are shown in the gray rows, and the WERs broken down by gender (female, male) are shown in the white rows.

(a) **Read** speech by TDNNF hybrid system

Country	NL				BE
Region	W	T	N	S	FL
DC/FC	N/A	23.8	28.3	25.6	35.3
	N/A	21.9,25.5	26.3,30.2	31.2,21.0	32.4,38.8
DT/FT	14.0	15.7	13.7	14.0	30.1
	12.7,15.0	13.2,17.7	12.8,14.6	12.6,15.5	28.6,31.8
DOA/FOA	17.2	19.0	13.3	25.0	22.5
	14.8,23.4	19.3,18.5	12.6,15.0	22.4,29.3	22.0,23.2

(b) **HMI** speech by TDNNF hybrid system

Country	NL				BE
Region	W	T	N	S	FL
DC/FC	N/A	31.4	27.0	30.1	47.5
	N/A	31.9,30.7	27.1,25.7	34.4,26.7	47.7,47.4
DT/FT	22.6	19.7	22.6	23.8	35.5
	19.1,25.8	19.2,19.9	23.1,21.6	23.4,23.9	34.6,36.7
DOA/FOA	29.0	29.3	24.3	37.4	36.4
	22.6,37.8	29.4,29.2	23.1,30.2	36.6,39.1	35.5,37.7

(c) **Read** speech by conformer E2E system

Country	NL				BE
Region	W	T	N	S	FL
DC/FC	N/A	26.5	30.9	27.7	39.1
	N/A	25.5,27.4	28.7,33.1	34.3,22.7	36.8,41.9
DT/FT	16.2	19.2	15.3	16.2	36.2
	14.2,17.9	16.0,22.3	14.2,16.3	14.8,17.9	33.3,39.5
DOA/FOA	18.7	20.2	14.7	26.9	24.3
	16.2,25.0	20.3,20.0	14.3,15.8	23.9,32.0	23.6,25.3

(d) **HMI** speech by conformer E2E system

Country	NL				BE
Region	W	T	N	S	FL
DC/FC	N/A	30.0	29.3	29.6	48.9
	N/A	29.7,30.4	28.2,30.9	32.4,27.4	48.5,49.3
DT/FT	20.4	19.9	22.7	25.0	39.8
	17.2,23.8	20.3,19.6	20.5,24.8	24.3,25.7	37.7,43.3
DOA/FOA	29.9	29.4	24.4	37.2	38.2
	24.1,37.9	29.4,29.3	22.4,33.7	35.1,41.4	37.9,38.6

B.3.5. WERs for read vs. HMI speech

Table B.3 shows that for both the hybrid system and the E2E system, the WER performance of HMI speech was much worse than that of read speech on native speaker groups. For the non-native speakers, the WER performances on read and HMI speech were very close for both ASR systems.

The tiny performance gap between non-native speakers' HMI and read speech indicates that the clarity of articulation is different in native and non-native speakers – native speakers tend to enunciate while reading out loud and tend to articulate less well during spontaneous (HMI) speech, while this articulation difference due to speaking style seems to be less for non-native speakers.

Appendix C. Character error rate details of the Mandarin ASRs

The CERs achieved by the TDNNF hybrid system and the conformer E2E system averaged over all speakers in the MagicData (adult-only) test set were 3.3% and 2.9% respectively. Table C.2 shows the CER for the female and male speech separately and averaged over both genders (column Avg), for every Mandarin accent region separately. The CERs per gender averaged over all the accent regions (row "All") are also shown.

Table C.2
CERs of the TDNNF hybrid system and the conformer E2E system on the MagicData test sets. "F/M" indicates female/male. "Avg" indicates the average over all speakers.

Arch.	Hybrid			E2E		
Set	F	M	Avg	F	M	Avg
All	3.1	3.5	3.3	2.7	3.1	2.9
NG	3.3	3.1	3.2	3.0	2.5	2.9
SG	2.9	2.8	2.8	2.3	2.5	2.5
[a]GA	2.4	2.3	2.4	2.1	2.3	2.2
[a]MI	3.8	6.4	5.1	3.4	5.6	4.5
[a]WU	1.9	3.0	2.6	1.5	2.8	2.3
[a]XI	3.2	3.5	3.4	2.9	2.9	2.9
[a]YU	3.3	3.8	3.5	2.7	3.8	3.2

[a] Regions use non-Mandarin Sinitic languages as their local languages.

C.1. CERs per gender

Table C.2 shows that, in general, female speech was better recognized than male speech. This is true for both the hybrid and the E2E architectures. This finding is in line with what has been found in the Dutch ASR experiments (see Appendix B.3.1).

Comparing the two ASR architectures shows that the E2E system outperformed the hybrid system on both female speech and male speech, both with an absolute CER reduction of 0.4%.

C.2. CERs per regional accent

Table C.2 shows that the region GA achieved the best recognition performance among the seven Chinese accent regions, and this is true for both the hybrid and the E2E ASR architectures. The two regions using a variety of Mandarin as local languages, i.e., NG and SG, achieved CER results that were on par with or lower than the average CER over all the regions. This also means overall, the regions using non-Mandarin Sinitic languages as their local languages have higher CER than regions using Mandarin. The region MI, in which the local language is not Mandarin, had the worst CER results by both the hybrid and the E2E ASR systems. Comparing the two ASR architectures shows that the E2E system was consistently better than the hybrid system on every accent region.

References

Abu Shariah, M., Sawalha, M., 2013. The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus. In: Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2.

Adda-Decker, M., Lamel, L., 2005. Do speech recognizers prefer female speakers? In: Proc. INTERSPEECH.

Alsharhan, E., Ramsay, A., 2020. Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition. Lang. Resour. Eval. 54 (4), 975–998.

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356 (6334), 183–186.

Chen, Z., Watanabe, S., Erdogan, H., Hershey, J.R., 2015. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In: Sixteenth Annual Conference of the International Speech Communication Association.

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.

Feng, S., Lee, T., 2018. Improving cross-lingual knowledge transferability using multilingual TDNN-BLSTM with language-dependent pre-final layer. In: Proc. INTERSPEECH. pp. 2439–2443.

Garnerin, M., Rossato, S., Besacier, L., 2019. Gender representation in French broadcast corpora and its impact on ASR performance. In: Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery. pp. 3–9.

Goldwater, S., Jurafsky, D., Manning, C.D., 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Commun. 52 (3), 181–200.

Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., Kamo, N., Li, C., Garcia-Romero, D., Shi, J., et al., 2021. Recent developments on espnet toolkit boosted by conformer. In: Proc. ICASSP. pp. 5874–5878.

Halpern, B.M., van Son, R., van den Brekel, M.W.M., Scharenborg, O., 2020. Detecting and analysing spontaneous oral cancer speech in the wild. In: Proc. INTERSPEECH. pp. 4826–4830.

Hori, T., Watanabe, S., Zhang, Y., Chan, W., 2017. Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. In: Proc. INTERSPEECH. pp. 949–953.

Hutiri, W.T., Ding, A.Y., 2022. Bias in automated speaker recognition. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 230–247.

Karita, S., Wang, X., Watanabe, S., Yoshimura, T., Zhang, W., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N.E.Y., Yamamoto, R., 2019. A comparative study on transformer vs RNN in speech applications. In: Proc. ASRU. pp. 449–456.

Kim, S., Hori, T., Watanabe, S., 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4835–4839.

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. Proc. Natl. Acad. Sci. 117 (14), 7684–7689.

Kudina, O., de Boer, B., 2021. Co-designing diagnosis: Towards a responsible integration of Machine Learning decision-support systems in medical diagnostics. J. Eval. Clin. Pract..

Leeuwen, D.A.v., Kessens, J., Sanders, E., Heuvel, H.v.d., 2009. Results of the N-best 2008 dutch speech recognition evaluation. In: Proc. INTERSPEECH.

Li, Q., Russell, M.J., 2001. Why is automatic recognition of children's speech difficult? In: Seventh European Conference on Speech Communication and Technology.

Magic Data Technology Co., Ltd., 2019. MAGICDATA Mandarin Chinese Read Speech Corpus. URL: http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101.

Moro-Velázquez, L., Cho, J., Watanabe, S., Hasegawa-Johnson, M.A., Scharenborg, O., Kim, H., Dehak, N., 2019. Study of the performance of automatic speech recognition systems in speakers with parkinson's disease. In: Proc. INTERSPEECH. pp. 3875–3879.

Oostdijk, N., 2000. The spoken dutch corpus. Overview and first evaluation. In: LREC. Athens, Greece, pp. 887–894.

Palanica, A., Thommandram, A., Lee, A., Li, M., Fossat, Y., 2019. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. NPJ Digit. Med. 2 (1), 1–6.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S., 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks. In: Proc. INTERSPEECH 2018. pp. 3743–3747.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: Proc. ASRU.

Qian, Y., Evanini, K., Wang, X., Lee, C.M., Mulholland, M., 2017. Bidirectional LSTM-RNN for improving automated assessment of non-native children's speech. In: INTERSPEECH. pp. 1417–1421.

Scharenborg, O., 2021. Inclusive Speech Technology: developing Automatic Speech Recognition for Everyone. Webinar delivered to the TU Delft Safety and Security Institute and Campus, The Hague, the Netherlands.

Schuster, M., Maier, A., Haderlein, T., Nkenke, E., Wohlleben, U., Rosanowski, F., Eysholdt, U., Nöth, E., 2006. Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition. Int. J. Pediatr. Otorhinolaryngol. 70 (10), 1741–1747.

Sun, S., Yeh, C.-F., Hwang, M.-Y., Ostendorf, M., Xie, L., 2018. Domain adversarial training for accented speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4854–4858.

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. pp. 53–59.

Tatman, R., Kasten, C., 2017. Effects of talker dialect, gender & race on accuracy of bing speech and YouTube automatic captions. In: Proc. INTERSPEECH. pp. 934–938.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., Ochiai, T., 2018. ESPnet: End-to-end speech processing toolkit. In: Proc. INTERSPEECH. pp. 2207–2211.

Wikipedia contributors, 2021. Jiangsu province. URL: https://zh.wikipedia.org/w/index.php?title=%E6%B1%9F%E8%8B%8F%E7%9C%81&oldid=66464253, [Online; accessed 21-July-2021].

Winata, G.I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., Fung, P., 2020. Learning fast adaptation on cross-accented speech recognition. In: Meng, H., Xu, B., Zheng, T.F. (Eds.), Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020. ISCA, pp. 1276–1280. http://dx.doi.org/10.21437/Interspeech.2020-0045.

Wu, Y., Rough, D., Bleakley, A., Edwards, J., Cooney, O., Doyle, P.R., Clark, L., Cowan, B.R., 2020. See what i'm saying? Comparing intelligent personal assistant use for native and non-native language speakers. In: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services. pp. 1–9.

Xu, H., Li, K., Wang, Y., Wang, J., Kang, S., Chen, X., Povey, D., Khudanpur, S., 2018. Neural network language modeling with letter-based features and importance sampling. In: Proc. ICASSP. pp. 6109–6113.

Yi, J., Wen, Z., Tao, J., Ni, H., Liu, B., 2018. CTC regularized model adaptation for improving LSTM RNN based multi-accent mandarin speech recognition. J. Signal Process. Syst. 90 (7), 985–997.

Zhang, Y., Zhang, Y., Halpern, B.M., Patel, T., Scharenborg, O., 2022. Mitigating bias against non-native accents. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Vol. 2022. pp. 3168–3172.

Zhang, Y., Zhang, Y., Patel, T., Scharenborg, O., 2022. Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems. In: Proc. 1st workshop on speech for social good (S4SG). pp. 15–19.

Zheng, H., Zhang, S., Qiao, L., Li, J., Liu, W., 2016. Improving large vocabulary accented mandarin speech recognition with attribute-based I-Vectors. In: Proc. INTERSPEECH. pp. 3454–3458.