Research Article

Yuanyuan Zhang*, Thomas De Valck and Odette Scharenborg

Speech recognition performance disparities between Dutch diverse speaker groups

https://doi.org/10.1515/phon-2025-0061
Received October 15, 2025; accepted March 18, 2026; published online April 22, 2026

Abstract: Current state-of-the-art automatic speech recognition (ASR) systems recognize typical speech (very) well. However, recent research has shown that their performance degrades for "diverse" speech, i.e., speech that diverges from "typical" speech due to, among others, demographic and sociolinguistic factors. In this work, given the rapid development of ASR technologies, we examined the performance of nine recently released ASR systems developed by Google, Microsoft, Meta, NVIDIA, and OpenAI, and three custom ASR models trained from scratch, on Dutch diverse speech. Our results showed that although overall recognition results differ quite substantially between the different systems, all systems show similar patterns regarding recognition performance for diverse speaker groups: for most ASR systems and models, language proficiency differences and severe speech motor impairment had a greater impact on performance disparities between speaker groups than demographic or sociolinguistic factors, indicating that acoustic variability due to demographic and sociolinguistic factors is well-represented in "typical speech" training data and consequently is well-modeled in the models. Furthermore, we found that differences in data processing pipelines and decoding setups significantly influenced recognition performance. Importantly, updates to company-developed ASR systems do not always improve performance of or reduce performance disparities between diverse speaker groups.

Keywords: performance disparities; automatic speech recognition; Dutch diverse speech; non-native accents; dysarthric speech

***Corresponding author: Yuanyuan Zhang**, Multimedia Computing Group, Delft University of Technology, Postbus 5031, 2600 GA, Delft, The Netherlands, E-mail: y.zhang-44@tudelft.nl. https://orcid.org/0009-0002-8351-8851

Thomas De Valck and Odette Scharenborg, Multimedia Computing Group, Delft University of Technology, Postbus 5031, 2600 GA, Delft, The Netherlands, E-mail: thomasdevalck12@gmail.com (T. De Valck), o.e.scharenborg@tudelft.nl (O. Scharenborg)

1 Introduction

Automatic speech recognition (ASR) has developed rapidly in recent years. For instance, the Microsoft ASR model supported only six languages in 2018 (Iancu 2019), while it supported forty-seven languages and language variations in 2025 (Microsoft 2025). ASR is widely used in diverse applications, including voice assistants (Davitaia 2025; Wienrich et al. 2021; Zhang et al. 2019), search engines (Luo et al. 2025), and health-related applications (Elhadad et al. 2025; Johnson et al. 2014; Latif et al. 2020). Given the crucial role spoken language plays in daily life, it is essential that ASR systems are able to recognize the large variability in speech, including that due to a speaker's voice, speech motor capabilities (e.g., dysarthric [a type of speech impairment] and non-dysarthric speakers), demographic (age, gender), language proficiency (children/age, non-native speakers), and sociolinguistic (regional) differences, which is referred to as "inclusive automatic speech recognition" (Scharenborg 2021).

However, recent experimental evidence shows that state-of-the-art ASR systems – both company-developed systems (Fuckner et al. 2023; Koenecke et al. 2020; Palanica et al. 2019; Raes et al. 2024; Roll and Graham 2025; Serditova et al. 2025; Weilinghoff 2025; Wu et al. 2020) and custom models trained by researchers themselves (Feng et al. 2021, 2024; Herygers et al. 2023) – do not perform equally well across all speakers. For instance, Koenecke et al. (2020) found that five company-developed systems developed by Amazon, Apple, Google, IBM, and Microsoft recognized white American speakers' speech more accurately than black American speakers' speech. Roll and Graham (2025) showed that ten company-developed systems from NVIDIA, Microsoft, and OpenAI exhibited substantial performance differences between native and non-native English speakers. Serditova et al. (2025) demonstrated that Rev AI's ASR system showed a lower performance for male speakers than female speakers when recognizing Newcastle English, while Weilinghoff (2025) showed that Whisper models (Radford et al. 2023) from OpenAI performed differently on Nigerian and Scottish English. Beyond English, ASR performance disparities have also been observed in Dutch and Flemish Dutch. Recent results for Dutch showed that state-of-the-art ASR systems exhibit performance disparities due to gender, age, and regional and non-native accents for a hybrid TDNNF-HMM model, an end-to-end (E2E) Conformer model (Feng et al. 2021, 2024; Herygers et al. 2023), and two company-developed systems (Fuckner et al. 2023): wav2vec 2.0 (Baevski et al. 2020) from Meta and Whisper (Radford et al. 2023) from OpenAI. Furthermore, Raes et al. (2024) demonstrated that Whisper models with different sizes (tiny, base, small, medium, and large) exhibit performance disparities based on speakers' (binary) gender (labels).

Company-developed ASR systems are continuously further developed and thus typically evaluated on English speech (Abouelenin et al. 2025; Artificial Analysis 2024; Koenecke et al. 2020, 2024; NVIDIA 2024; Palanica et al. 2019; Roll and Graham 2025; Weilinghoff 2025; Wu et al. 2020). An open question is how well these systems perform on a non-English language and speech that deviates from typical speech, which we refer to as "diverse" speech (Zhang et al. 2023a). We address this question by evaluating the recognition performances of a variety of company-developed ASR systems and custom models on Dutch diverse speech from speakers that cover a wide range of acoustic variability. Following Koenecke et al. (2020), we evaluate Google Cloud systems, including Chirp 2 (Zhang et al. 2023b) and Google Telephony (Google Cloud 2025), the Microsoft Azure ASR system (Microsoft 2025), and Meta's Massive Multilingual Speech system (Pratap et al. 2024). Additionally, we evaluate NVIDIA's NeMo-nl system (NVIDIA 2024), and three versions of OpenAI's Whisper (Radford et al. 2023): Whisper-large-V2, Whisper-large-V3, and Whisper-large-V3-turbo.

Furthermore, we trained three custom Dutch ASR models from scratch using publicly available Dutch corpora and compared their performance to the above-mentioned systems. Training custom models aligns with the broader goal of developing fully reproducible ASR models (Peng et al. 2023, 2024, 2025) that are trained with transparent and reproducible training data, in contrast to and as a reaction to several company-developed systems for which the ASR architectures and training data are unknown (e.g., Google Chirp 2, Google Telephony, and Microsoft Azure systems, while the training data of Whisper is unknown). Moreover, company-developed ASR systems are typically trained on large amounts of speech data, which leads to high computational costs, which in current times of climate change and energy crises are hard to justify (Parcollet and Ravanelli 2021). Following Feng et al. (2024), we trained three custom models on a much smaller, Dutch speech database, the Corpus Gesproken Nederlands (Oostdijk 2000). Wav2vec 2.0 features (Baevski et al. 2020) have achieved state-of-the-art performance on both typical (Chang et al. 2021) and atypical speech recognition tasks (Hernandez et al. 2022; Sapkota et al. 2025). Whisper features were found to be robust to diverse downstream tasks, e.g., keyword spotting (Chemudupati et al. 2023). To train the best custom models, we therefore investigated the use of large-scale pre-trained features. Specifically, we trained Dutch ASR models using acoustic features extracted from Whisper-large-V2 and Wav2vec 2.0's XLSR-53 (Conneau et al. 2021) and compared these to a baseline model trained with conventional filter bank (FBank) features (Davis and Mermelstein 1980).

We tested the effect of demographic (age, gender), language proficiency (children/age, non-native speakers), and sociolinguistic (regional) differences on the recognition performance of the company-developed ASR systems and the three custom models on two varieties of Dutch: Dutch as spoken in the Netherlands and in Flanders,

i.e., Flemish, from the Jasmin corpus (Cucchiarini et al. 2006). To investigate the effect of speech motor capabilities, we tested all systems and models on the dysarthric speech from a native Dutch male speaker from the DysOne dataset (Zhang et al. 2026), since Jasmin does not contain this type of diverse speech. Finally, we investigated the effect of system settings of some company-developed systems on recognition performance. To our knowledge, this work presents the first comprehensive evaluation of both company-developed state-of-the-art ASR systems and models trained with large pre-trained features across Dutch diverse speaker groups, and as such will be an important milestone and benchmark for the development of inclusive ASR for Dutch.

2 Methodology

2.1 Datasets

2.1.1 Jasmin

Jasmin contains 29.9 h of Dutch read speech and 10.6 h of human-machine interaction (HMI) speech, 17.9 h of Flemish read speech and 7.2 h of HMI speech. Both Dutch and Flemish speech was produced by speakers from five speaker groups with binary gender labels: NC: native children, NT: native teenagers, NOA: native older adults, NNC: non-native children, and NNA: non-native adults. For the native speakers, regional information is provided in the corpus. Dutch native speakers are from four accent regions: core (C), transitional (T), northern peripheral (NP), and southern peripheral (SP). Note, Jasmin does not contain speech from Dutch native child speakers from the core region. The Flemish native speakers are grouped into: core (C), transitional (T), west peripheral (WP), and Limburg peripheral (LP). Typically, each speaker provided read and HMI speech. For the non-native speakers of Dutch and Flemish, common European framework of reference for languages (CEFR) proficiency information was provided for the adult speakers. The most proficient non-native speaker had a level of B1, while most had level A2 or A1 – note, no proficiency information was provided for child speakers. Table 1 shows the number of native speakers broken down by gender (female, male) for each age group and each region. Table 2 shows the number of non-native speakers broken down by gender for each age group and each proficiency level in the Netherlands and Flanders.[1]

Following the data processing pipeline in Feng et al. (2024), the recordings were segmented into utterance-based chunks, and silences were removed according to the

1 The demographic table regarding regional information was obtained using https://metaspeech.ewi.tudelft.nl/ to analyze the speaker information provided in the Jasmin corpus.

Table 1: Number of native female and male speakers (x, y) of Dutch and Flemish per age group per region in the Jasmin corpus.

Age group	Dutch regions				Flemish regions			
	C	T	NP	SP	C	T	WP	LP
NC	0, 0	15, 14	11,12	9, 11	5, 4	6, 4	6, 6	6, 6
NT	9, 11	2, 2	10, 10	10, 9	5, 3	5, 3	8, 7	5, 4
NOA	13, 5	9, 8	13, 4	10, 6	5, 3	5, 3	7, 5	5, 4

Table 2: Numbers of non-native female and male speakers (x, y) of Dutch and Flemish per age group per proficiency level in the Jasmin corpus.

Age group	Dutch speakers				Flemish speakers			
	B1	A2	A1	Unknown	B1	A2	A1	Unknown
NNC	–	–	–	28, 25	–	–	–	25,27
NNA	6, 3	18, 8	4, 6	–	8, 3	5, 4	6, 3	0, 1

time stamps provided by the Jasmin-CGN corpus. The duration range of the resulting segments is 0.2–14.4 s. The machine-generated speech in the HMI dialogues, which was saved in one of the two recording channels, was removed, as were utterances containing only non-linguistic content.

2.1.2 DysOne

The DysOne dataset is a bilingual (native Dutch and non-native English) and bimodal (speech and video) dataset containing recordings of a 35-year-old male native Dutch speaker with severe dysarthria (Zhang et al. 2026). For both Dutch and non-native English, DysOne contains two speech types: read and spontaneous speech. In this study, we used the Dutch part of DysOne consisting of 3.3 h of read speech and 0.4 h of spontaneous speech.

2.2 ASR systems and models

Table 3 shows an overview of the nine company-developed ASR systems and the three custom models trained by us that were used in this study, indicates for each system and model whether the training data is known (Trans [parent]), and whether

Table 3: Overview of the 12 ASR systems and models evaluated in this work. Trans denotes transparent training data. Multi denotes multilingual systems/models.

	GC	GT	MNL	MVL	MMS	WV2	WV3	WV3T	NM	CF	CWV2	CX
API	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
Trans	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
Multi	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗

it is a multilingual (Multi) or Dutch-only system/model, since these aspects affect their accessibility, reproducibility, and performance.

Four of the nine ASR systems have commercial API systems: Google Chirp 2 (GC), Google Telephony (GT), and the Microsoft Azure systems for Dutch (MNL) and for Flemish (MVL). The other five company-developed systems were downloaded from Hugging Face: Meta's Massively Multilingual Speech (MMS), three versions of Whisper of OpenAI: Whisper-large-V2 (WV2), Whisper-large-V3 (WV3), and Whisper-V3-Turbo (WV3T), and NVIDIA's NeMo-nl (NM). Additionally, we trained three custom Conformer ASR models from scratch with three types of acoustic features, two of which were extracted from company-developed models: FBank (Davis and Mermelstein 1980), Whisper-large-V2 (Radford et al. 2023), and Wav2Vec 2.0's XLSR-53 (Conneau et al. 2021), referred to as CF, CWV2, and CX, respectively.

2.2.1 The company-developed systems

Google Chirp 2 (GC) is the latest version of the Chirp model (Zhang et al. 2023b) as of August 2025. Since our test data includes both Dutch and Flemish speech, we initially aimed to use two Google systems, with one specifically for Dutch and the other for Flemish. However, Google Chirp 2 only contains the Dutch language setting. Since Google Telephony (GT) has the language setting for Flemish, we used Google Telephony in addition to Google Chirp 2. Google Telephony is optimized for telephone speech (Google Cloud 2025). For both Google systems, we performed synchronous speech recognition using the Google Cloud API via the 2.33.0 version of Speech-to-text V2 python package,[2] where the entire audio input is processed before returning results. All utterances in our test data are shorter than 30s; therefore, no speech data

2 We also evaluated the Google Speech-to-text V1 systems in June 2024 by setting the language to Dutch and Flemish, respectively. However, their performance was substantially worse than those of Google Chirp 2 and Google Telephony. Therefore, in this work, we only discuss the results of the newest release.

from our test sets were stored on the Google Cloud. The Google Cloud API provides free credits.

We employed the Microsoft Azure ASR system using two language settings, i.e., Dutch (MNL) and Flemish (MVL).[3] For both Microsoft systems, we used synchronous recognition. We used Microsoft Azure API via the 1.45.0 version of Azure cognitive services speech python package. The Microsoft Azure API provides free credits.

We downloaded the Massive Multilingual Speech (MMS) model "mms-1b-all" (Meta AI 2023; Pratap et al. 2024) by Meta from Hugging Face for our evaluation. MMS was built upon the framework of wav2vec 2.0 (Baevski et al. 2020). It was first pre-trained with 491 k hours of speech in 1,406 languages and then fine-tuned on speech from 1,162 languages (Pratap et al. 2024).

Similarly, three Whisper models (WV2, WV3, and WV3T) were downloaded from Hugging Face. WV2 was trained on 680 k hours of multilingual speech, mostly in English (Radford et al. 2023). WV3 was trained on even more training data than WV2, achieving better ASR performance across languages and accents (OpenAI 2023a). WV3T achieves comparable performance with WV3 but is optimized for faster inference (OpenAI 2023b). During testing, we employed beam search decoding with a beam size of 10; we set the task to "transcribe", the language to "Dutch", and the temperature parameter to 0.

NeMo-nl (NM) is an ASR system for Dutch developed by NVIDIA. NM was downloaded from Hugging Face. NM is trained on a combination of three publicly available Dutch datasets, containing 621 h of speech data (NVIDIA 2023, 2024): (1) Common Voice 12 (40 h) (Ardila et al. 2019), consisting of read speech from volunteers online; (2) Multilingual LibriSpeech (547 h) (Pratap et al. 2020), consisting of read speech from audio books; and (3) VoxPopuli (34 h) (Wang et al. 2021), consisting of semi-spontaneous parliamentary debates. NM is trained primarily on Dutch, with any Flemish representation being incidental (e.g., via European Parliament from VoxPopuli or crowd sourced volunteers from Common Voice). During decoding, we used beam search with a beam size of 10.

2.2.2 Custom ASR models

Following Feng et al. (2021, 2024), we trained Conformer-based ASR models on the Corpus Gesproken Nederlands (CGN; Spoken Dutch Corpus) (Oostdijk 2000), using 80-dimensional FBank features (CF), 1280-dimensional features extracted from the

3 We tested these two systems at two time points, in June 2024 and July 2025. Here, we present the results of the Microsoft 2025 systems, with comparisons to the results of the 2024 systems described in Section 3.3.2.

encoder of WV2 (CWV2), and 1024-dimensional fused representations from all 24 layers of XLSR-53, following the feature extraction method used in Hernandez et al. (2022) (CX). CGN contains both Dutch and Flemish spoken by native adult speakers from the Netherlands and Flanders. CGN contains approximately 900 h of raw speech recorded in a wide range of recording settings including lectures, read speech, telephone conversations (CTS), and broadcast news (BN). The entire CGN dataset, including both language varieties, was used for training. After segmenting the recordings into smaller chunks, removing silent segments, and excluding utterances shorter than 0.1 s,[4] 690.5 h (Dutch: 424.6 h; Flemish: 265.9 h) were used as the training set, and 6.9 h were used as a validation set.

We employed a medium-sized Conformer encoder (Gulati et al. 2020) with a Transformer decoder. The model consists of 12 encoder layers and 6 decoder layers. Each encoder layer has a model dimension of 256, a feed-forward dimension of 1,024, and uses 4 attention heads. The decoder uses a feed-forward dimension of 2,048 and has 4 attention heads. The encoder begins with a 2-layer convolutional subsampling module with 256 channels, kernel size 3, and stride 2. Each Conformer block includes a convolutional module with the default kernel size of 31. The Conformer model was trained using a joint connectionist temporal classification (CTC)-attention objective (Kim et al. 2017), with a CTC weight of 0.3, and an attention weight of 0.7. During training, we applied two-fold speed perturbation with factors of 0.9 and 1.1, and SpecAugment (Park et al. 2019). Model training was conducted using the ESPnet toolkit (Watanabe et al. 2018). BPE units were set to 5,000, and the batch bin size was set to 16 K. Training was performed for up to 50 epochs, with early stopping (patience = 3) based on validation loss. For inference, the final ASR model was obtained by averaging the parameters of the ten checkpoints with the lowest validation loss. During decoding, we used beam search with a beam size of 10.

The models were validated on four in-domain CGN test sets, following the split in van Leeuwen et al. (2009): a Dutch CTS test set (1.8 h), a Dutch BN test set (0.4 h), a Flemish CTS test set (1.7 h), and a Flemish BN test set (0.9 h). Their performance on the CGN test sets is: for the CF model, the WERs on NL-BN, NL-CTS, FL-BN, and FL-CTS are 5.9 %, 19.5 %, 6.7 %, and 22.8 %, respectively; for CWV2, the WERs are 7.0 %, 27.6 %, 8.8 %, and 31.9 %, respectively; for CX, the WERs are 5.7 %, 21.8 %, 6.9 %, and 25.7 %, respectively. All three models achieved state-of-the-art performance on the CGN test sets compared to results reported in (Feng et al. 2024; Patel and Scharenborg 2024; Zhang et al. 2023a).

4 Utterances shorter than 0.1s cause errors when extracting acoustic features using Whisper-large-V2 and XLSR-53 models.

2.3 Evaluation

To fairly compare the decoding results of the different ASR systems and models, we performed text normalization on both the ground truth transcriptions and the systems/models' results. Specifically, we removed all punctuation except the apostrophe, since it is linguistically important in Dutch. All characters were converted to lowercase, and Arabic numerals were transcribed into their Dutch textual forms.

We measured speech recognition performance in Word Error Rates (WERs) and report insertions, substitutions, and deletions. We used the sclite scoring tool (Fiscus 2015) to compute the WERs. When calculating the WERs, the non-linguistic symbols in the ground truth transcripts were ignored. In addition to WERs, we report the performance disparities between different speaker groups, defined as the WER difference between two speaker groups (Feng et al. 2024).

We conducted pairwise comparisons of the estimated marginal means (emmeans) (Lenth 2023) derived from generalized linear models (GLMs) (Nelder and Wedderburn 1972) in the R language (R Core Team 2021) to determine whether the performance disparities between different age groups (children, teenagers, older adults, and adults), different genders (female and male speakers), native and non-native speakers, and different regional accents (speakers from C, T, NP/WP, SP/WP regions), are statistically significant. We employed a gamma distribution for the GLMs because WERs are continuous, strictly non-negative, and typically exhibit a positively skewed distribution with variance that increases with the mean.

Since we want to investigate how each system/model performs on diverse Dutch speech, separate GLM models were constructed for each ASR system/model, for each language variety (Dutch or Flemish), and each speech type (read or HMI). When evaluating the effect of age, gender, and nativeness, we employed the following GLM function:

$$\text{WER per speaker} \sim \text{Gender X GroupID}$$

where Group IDs contain the five speaker groups of NC, NT, NOA, NNC, and NNA (see Section 2.1.1). We used group IDs instead of nativeness and age factors due to the structural sparsity of the data, i.e., certain demographic combinations do not exist (e.g., the non-native speaker groups lack teenagers and older adults). When evaluating the effect of regional accents, we employed the following GLM function:

$$\text{WER per speaker} \sim \text{AgeRegionID}$$

using only native speakers' results, where AgeRegionID is a joint factor of three age groups for native speakers (children, teenagers, and older adults) and different regions (C, T, NP/WP, and SP/WP). We used the joint factor instead of age and region

factors, again, due to the structural sparsity of the data, i.e., there are no Dutch children from the core region (see Table 1).

To verify the validity of our GLMs, we employed the performance package (Lüdecke et al. 2021) in R for diagnostics. All models successfully passed the assessments for posterior predictive fit, collinearity, influential outliers, and the homogeneity of standardized residuals. While five models (pertaining to age, gender, and nativeness) exhibited slight underdispersion (dispersion ratios of 0.74–0.78), this remained within an acceptable range. One model for regional accent showed overdispersion. However, as this model yielded only one significant comparison, the impact on the overall results is minimal.

Since we conducted multiple comparisons, to account for the risk of Type I errors (false positives), the resulting p-values were corrected using the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg 1995). This correction was applied independently to all comparisons for each factor (age, gender, nativeness, and regional accent) within every GLM model. For each comparison, we report the corresponding p-value and effect size measured by Cohen's d (Cohen 1988). We did not perform statistical testing for dysarthric speech, as the DysOne dataset contains only a single speaker.

3 Results and analyses

3.1 Overall recognition performance of the different systems and models

We first compared the recognition performance of the different systems and models, evaluating them on both non-dysarthric and dysarthric speech (Jasmin and DysOne), reporting WERs, insertions, deletions, and substitutions, and the effect of sentence length on performance. For Jasmin, results were reported for Dutch and Flemish read and HMI speech, separately for native and non-native speakers, and the average across these two groups. For DysOne, results were reported for read and spontaneous speech, separately.

3.1.1 Overall performance on non-dysarthric speech

Table 4 shows the WER of all 12 systems and models, split for Dutch and Flemish speech, read and HMI speech, and split for native and non-native speakers. Google

Table 4: WERs (%) of the ASR systems and models on the Jasmin corpus. N = native speakers, NN = non-native speakers, Avg = average over N and NN. Bold indicates the lowest WER among systems and models for the same test set.

Model	Dutch read			Dutch HMI			Flemish read			Flemish HMI		
	N	NN	Avg	N	NN	Avg	N	NN	Avg	N	NN	Avg
GC	**12.4**	**27.6**	**17.5**	23.5	**32.5**	**26.5**	**12.2**	**19.2**	**14.9**	21.3	**24.3**	**22.5**
GT	12.9	28.0	17.9	**22.2**	37.6	27.4	13.4	20.2	16.0	**21.2**	32.6	25.8
MNL	17.4	32.2	22.3	26.1	35.1	29.2	16.9	24.2	19.7	24.8	28.8	26.4
MVL	15.0	30.3	20.1	23.8	34.0	27.3	13.8	20.8	16.4	23.2	28.4	25.3
MMS	38.3	70.9	49.1	58.3	80.7	65.9	36.9	55.6	44.0	53.8	74.8	62.2
WV2	20.7	37.6	26.3	33.2	45.8	37.4	18.9	29.5	23.0	29.9	34.8	31.9
WV3	14.5	31.1	20.1	27.9	39.0	31.7	14.2	22.9	17.5	25.1	29.9	27.0
WV3T	17.9	40.2	25.3	31.9	46.3	36.7	19.3	29.6	23.3	32.1	36.9	34.1
NM	41.1	69.0	50.4	56.4	69.0	60.6	37.4	51.1	42.6	52.3	55.9	53.7
CF	24.3	49.5	32.7	34.7	54.3	41.3	16.7	32.2	22.6	23.5	35.9	28.5
CWV2	26.5	44.5	32.5	39.1	53.9	44.1	22.9	34.9	27.5	31.0	39.1	34.3
CX	18.4	42.8	26.5	27.6	46.0	33.8	14.8	27.6	19.7	21.8	31.4	25.6

Chirp 2 (GC) achieved the lowest WERs (i.e., best performance) on all four Jasmin test sets of non-dysarthric speech (Dutch read, Dutch HMI, Flemish read, and Flemish HMI; columns Avg). Most systems claim to have very good results on English (Artificial Analysis 2024; Radford et al. 2023) .[5] For instance, Google Chirp 2, Whisper-large-V3, and Microsoft Azure ASR systems achieved WERs of 6.8 %, 7.6 %, and 7.9 % on VoxPopuli data (semi-spontaneous speech; including non-native English speakers), respectively (Artificial Analysis 2024). We however observed large differences across the systems in the performances for Dutch and Flemish non-dysarthric speech (columns N and NN of Table 4). Massive Multilingual Speech (MMS) and NeMo-nl (NM) performed the worst on non-dysarthric speech. Table 5 shows the number of substitutions, deletions, and insertions of all 12 systems and models on all four Jasmin test sets. For almost all systems and models, the most occurring errors are substitutions followed by deletions.

Averaged over native and non-native speakers, each system performed better on Flemish than Dutch speech. This finding can be explained by the comparison of the

5 Excluding NeMo-nl which is a Dutch model and Massive Multilingual Speech whose aim is to work for a lot of different languages, and consequently their performance on English is not very good.

Table 5: Substitution (S), deletion (D), and insertion (I) error rates (%) of the ASR systems and models split for the four test sets of the Jasmin corpus.

Model	Dutch read			Dutch HMI			Flemish read			Flemish HMI		
	S	D	I	S	D	I	S	D	I	S	D	I
GC	12.0	4.0	1.5	16.8	6.7	3.0	10.5	3.0	1.4	14.3	5.6	2.6
GT	10.9	5.4	1.7	13.6	8.0	5.9	9.1	5.7	1.2	11.3	6.2	8.3
MNL	12.5	8.3	1.6	15.6	10.9	2.7	12.1	6.1	1.6	14.8	8.9	2.7
MVL	13.7	3.9	2.5	16.3	7.6	3.3	11.5	3.0	2.0	15.0	6.9	3.3
MMS	33.2	12.8	3.1	42.3	19.3	4.3	30.4	9.9	3.7	41.1	14.4	6.7
WV2	16.3	4.6	5.5	22.5	7.1	7.8	15.1	3.6	4.3	19.0	6.8	6.1
WV3	13.5	4.0	2.6	19.5	7.1	5.1	12.2	3.1	2.2	16.2	6.8	4.0
WV3T	17.4	4.7	3.2	23.6	7.7	5.5	15.1	5.6	2.6	20.4	9.3	4.3
NM	13.3	35.3	1.8	19.0	38.5	3.1	15.0	25.9	1.7	19.2	31.6	3.0
CF	12.1	17.9	2.7	12.8	24.4	4.1	11.1	8.4	3.2	11.3	12.8	4.3
CWV2	17.0	11.8	3.7	17.6	21.2	5.3	14.6	9.1	3.8	14.4	15.0	4.9
CX	13.5	10.0	3.0	14.6	14.6	4.6	10.9	5.8	3.0	11.4	10.2	4.0

performances on the native and non-native speech separately which shows that non-native accented Dutch is much worse recognized that non-native accented Flemish (see NN columns). Further investigation of the recognition performances for the non-native speakers split per proficiency level showed that recognition performance for the Flemish non-native speech was consistently much better than that for the Dutch non-native speech (ranging from ~6 % to 15 % WER difference). Potentially, the much better performance for the Flemish non-native speech is due to speaker characteristics, sentence length differences, or linguistic content rather than proficiency level, as these were fairly similar for the non-native Dutch and Flemish speakers. These results show the importance of examining ASR performance across diverse speaker groups, rather than focusing on an overall average.

Potentially, the differences in WER among different systems and models can be explained by differences in how they perform for different sentence lengths. Figure 1 shows the WERs by sentence length, measured in number of words per sentence. Across the four non-dysarthric speech test sets, sentences shorter than 5 words led to higher WERs for all systems and models. As sentence length increased, WERs decreased before rising again at around 15 words. Also, there is a striking similarity in the effect of sentence length on recognition performance, with one exception:

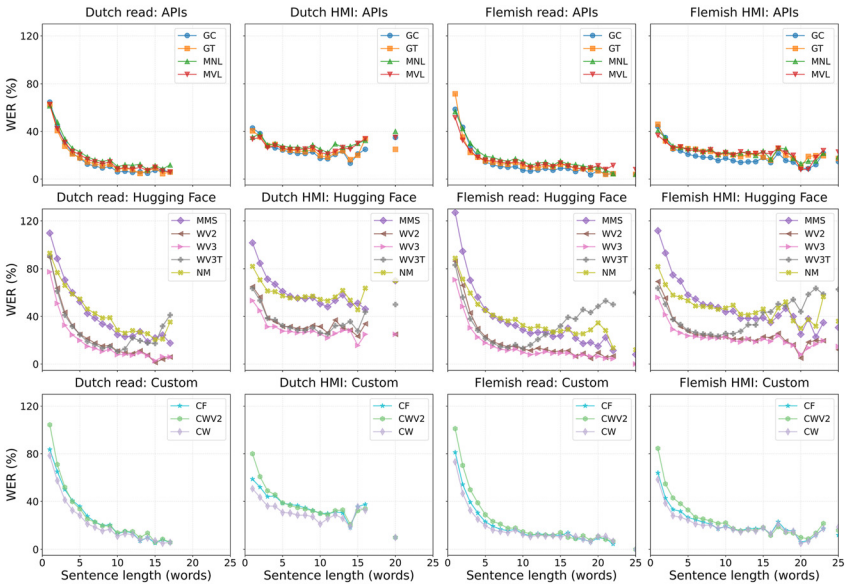

Figure 1: WERs for the different sentence lengths (measured in number of words per sentence) of the Dutch and Flemish read and HMI speech from the Jasmin corpus split for the four ASR systems employed via APIs (top row); the five ASR systems downloaded from Hugging Face (middle row), and the three custom Conformers (bottom row).

Whisper-large-V3-turbo (WV3T) exhibited this rise at much shorter lengths than other models, especially for Flemish, where WV3T's WER increased sharply after 10 words for read and HMI speech. This trend is not observed for Whisper-large-V3 (WV3), which suggests that architecture compression (as described in Section 2.2.1) might cause problems with longer sentences for WV3T.

3.1.2 Overall performance on dysarthric speech

Table 6 shows the WER and number of insertions, deletions, and substitutions of the 12 ASR systems and models on the dysarthric speech of DysOne. Overall, for our Dutch dysarthric speaker, all 12 systems and models performed very poorly, with WERs far higher than those for non-dysarthric speech, exhibiting substantial performance disparities between non-dysarthric speech and the severe dysarthric

Table 6: WERs, substitution (S), deletion (D), and insertion (I) error rates (%) of the ASR systems and models on the DysOne dataset. Bold indicates the lowest WER among systems and models on the same test set.

Model	Dysarthric read				Dysarthric spontaneous			
	WER	S	D	I	WER	S	D	I
GC	179.7	51.3	16.4	112.0	307.8	56.5	17.6	233.7
GT	89.0	13.8	74.7	0.5	97.0	7.3	89.1	0.6
MNL	91.6	24.7	64.6	2.3	96.8	8.9	87.5	0.4
MVL	93.0	23.2	69.3	0.6	96.7	10.5	86.2	0.0
MMS	99.5	60.8	37.9	0.9	99.5	65.3	32.8	1.5
WV2	103.5	61.6	7.9	34.0	151.4	62.5	9.9	78.9
WV3	**75.0**	55.3	8.2	11.5	**80.1**	59.8	8.7	11.6
WV3T	80.0	56.0	16.8	7.2	85.4	52.6	24.5	8.4
NM	98.7	3.5	95.2	0.0	99.2	5.1	94.0	0.1
CF	93.6	11.9	81.3	0.4	97.0	6.9	90.0	0.1
CWV2	75.3	32.9	39.0	3.5	82.8	24.7	55.1	3.0
CX	84.8	29.2	53.7	1.9	91.4	21.7	68.6	1.1

speech from DysOne. Whisper-large-V3 (WV3) performed the best, while Google Chirp 2 (GC) performed the worst on the dysarthric speech of DysOne.

The most occurring error depends heavily on the system or model: sometimes insertions, for other systems and models, deletions occurred most frequently, with two exceptions: the insertion rate for GC is exceptionally high, followed by that of Whisper-large-V2 (WV2), indicating that GC and WV2 hallucinated when recognizing dysarthric speech. Inspection of these insertions showed that these were so-called "hallucinations" (Ji et al. 2023), i.e., recognized output "that is nonsensical, or unfaithful to the provided source input" (Ji et al. 2023: 3). Nevertheless, overall, we thus do not observe many hallucinations for the systems we investigated. In Koenecke et al. (2024), when recognizing aphasia speech (a type of language disorder), hallucinations were also observed in Whisper API.

Unlike non-dysarthric speech, for the dysarthric speech, sentence length has little effect on recognition performance: the curves shown in Figure 2 for most systems and models (10 out of 12, excluding Google Chirp 2 (GC) and Whisper-large-V2 [WV2]) are relatively flat. GC and WV2 have high insertion rates (hallucinations) regardless of sentence length.

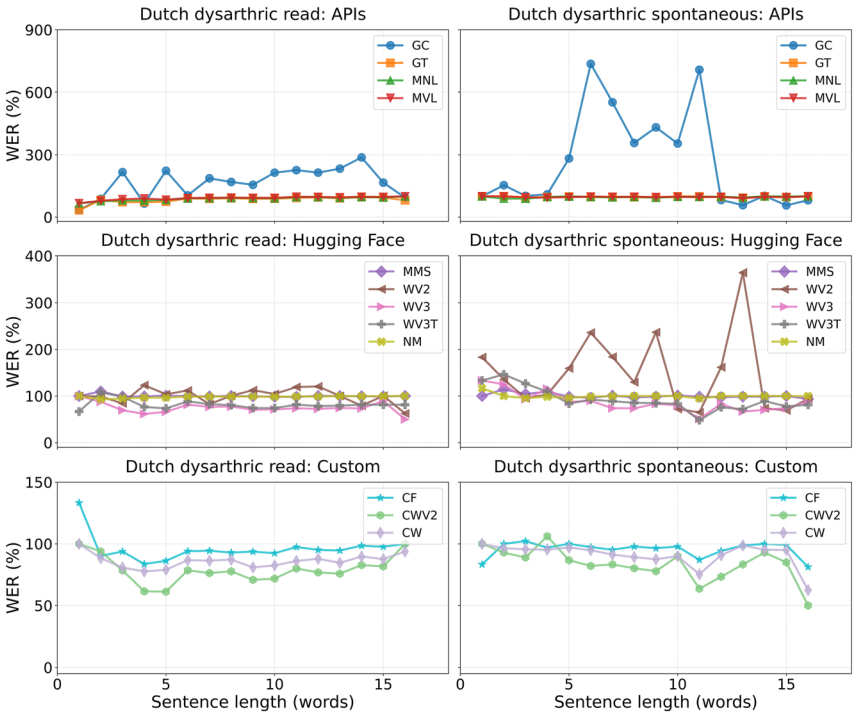

Figure 2: WERs for the different sentence lengths of Dutch dysarthric read and spontaneous speech from DysOne, split for the four ASR systems employed via APIs (top row), the five ASR systems downloaded from Hugging Face (middle row), and the three custom Conformers (bottom row).

3.1.3 The effect of speaking style

When comparing the effect of speaking style on recognition performance for non-dysarthric speech (see Table 4), all systems and models performed better on read speech than on more spontaneous speech, with lower substitution, deletion, and insertion error rates. This is in line with results in the literature, which consistently shows that for most systems and models, non-dysarthric read speech is more intelligible than spontaneous speech (Feng et al. 2024; Fuckner et al. 2023; Russell et al. 2024).

When comparing the effect of speaking style on recognition performance for dysarthric speech (see Table 6), almost all systems and models performed better on read speech than on more spontaneous speech, with lower substitution, deletion, and insertion error rates. Only Massive Multilingual Speech (MMS) and NeMo-nl (NM) obtained similar WERs on dysarthric read and spontaneous speech. For dysarthric

speech, in general, read speech showed higher substitution but lower deletion rates compared to spontaneous speech. We have not been able to find any studies comparing ASR performance on dysarthric read and spontaneous speech; although, a phonetic analysis of the speech of speakers with severe dysarthria did not show more anomalies in spontaneous speech than in read speech (Laaridh et al. 2016).

3.2 The effect of demographic, language proficiency, and sociolinguistic factors on recognition performance

Figure 3 presents the WERs of all ASR systems and models split for the five different Dutch and Flemish speaker groups (NC, NT, NOA, NNC, and NNA). Overall, the company-developed API systems – Google Chirp 2 (GC), Google Telephony (GT), the Microsoft ASR system for Dutch (MNL), and the Microsoft ASR system for Flemish (MVL) – and the Whisper-large-V3 (WV3) consistently achieve lower error rates for each group of speakers, while the Massive Multilingual Speech model (MMS) and NeMo-nl (NM) show the worst performance. Despite being trained on only ~700 h of training data, the custom Conformer model trained with XLSR-53 features (CX) shows comparable performance to Whisper-large-V2 (WV2) and even outperforms all systems on native Flemish teenagers' HMI speech.

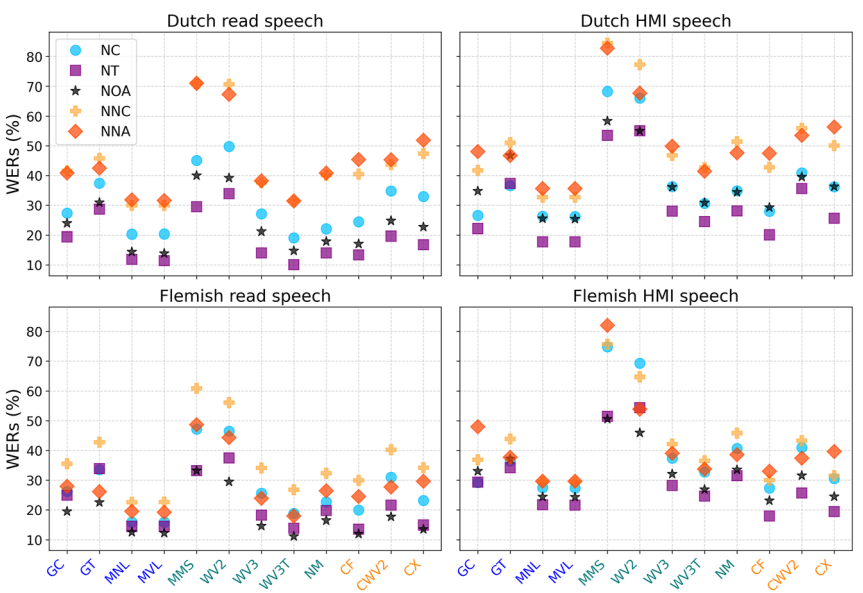

Figure 3: WERs of the ASR systems and models split for the five speaker groups.

3.2.1 The effect of age on recognition performance

Overall, for Dutch native speakers (top row in Figure 3), for both read and HMI speech, all systems and models performed the best on teenagers' speech (purple markers), followed by older adults' speech (gray markers), while performing the worst on children's speech (light-blue markers). For Dutch non-native speakers, for both read and HMI speech, most systems and models performed similarly on children's speech (yellow markers) and adults' speech (orange markers).

For Flemish native speakers (bottom row in Figure 3), similar to Dutch, for read speech, all systems and models performed the worst on children's speech. For HMI speech, Google Telephony (GT) performed similarly across the three age groups of speakers (children, teens, and older adults), while the other 11 systems and models performed the worst on children's speech. For Flemish non-native speakers, unlike in Dutch, for read speech, most systems and models performed better on adults' speech than that of children. For HMI speech, some systems and models performed better on children's speech than on adults', while others performed better on adults' speech.

Table 7 presents the size of the performance disparity between the two age groups listed in the first column for each system or model for each of the five speaker groups, split for Dutch (Table 7a and b) and Flemish (Table 7c and d) and read and HMI speech. For native Dutch speakers, we indeed observed substantial and significant effects of age of the speaker group on recognition performance for children in both read and HMI speech and for older adults in read speech across all systems and models. While most systems and models also showed a significant age effect for older adults in HMI speech, this was not observed for NeMo-nl (NM). For Dutch non-native speakers (NNC, NNA), in HMI speech, only NeMo-nl (NM) showed a significant age effect for child speech. No age performance disparities were observed between non-native adults and children for the other 11 systems and models, consistent with the lack of performance disparities found for the two Dutch ASR models trained from scratch as quantified in Feng et al. (2024). For Flemish native speakers, similar to Dutch, nearly all systems and models showed substantial age effects for child speech, with the exception of Microsoft Azure's ASR system for Flemish (MVL) which did not show an age effect for read speech, and Google Telephony (GT) and Microsoft Azure's ASR system for Dutch (MNL) which did not show an age effect for HMI speech. For Flemish non-native speakers, most systems and models exhibited an effect of age on recognition performance for either children or adults, with the exception of Microsoft Azure's ASR system for Flemish (MVL).

Overall, for both Dutch and Flemish, most systems and models exhibited substantial age performance disparities, particularly against children and, to a lesser extent, older adults, while patterns for non-native speakers were more variable, with

Table 7: Performance disparities between different age groups computed by subtracting the WER of the second mentioned age group in column 1 from that of the first mentioned age group. A positive WER difference indicates a lower WER for the second mentioned speaker group and vice versa for a negative WER difference. Bold denotes the performance disparity is significant. Statistical significance levels: *$p < 0.05$; †$p < 0.01$; ‡$p < 0.001$. *d* denotes the effect size.

	GC	GT	MNL	MVL	MMS	WV2	WV3	WV3T	NM	CF	CWV2	CX
(a) Dutch read												
NC, NT	**9.1‡ (*d* = 0.20)**	**8.0‡ (*d* = 0.20)**	**11.3‡ (*d* = 0.33)**	**8.7‡ (*d* = 0.23)**	**15.4‡ (*d* = 0.60)**	**12.8‡ (*d* = 0.31)**	**8.8‡ (*d* = −0.22)**	**8.1‡ (*d* = 0.22)**	**15.8‡ (*d* = 0.59)**	**16.1‡ (*d* = 0.43)**	**15.2‡ (*d* = 0.50)**	**11.0‡ (*d* = 0.31)**
NC, NOA	**6.0‡ (*d* = 0.13)**	**5.4‡ (*d* = 0.13)**	**8.8‡ (*d* = 0.25)**	**6.2‡ (*d* = 0.16)**	**5.1† (*d* = 0.24)**	**5.8† (*d* = 0.13)**	**4.2† (*d* = 0.10)**	**4.2† (*d* = 0.11)**	**10.5‡ (*d* = 0.39)**	**10.2‡ (*d* = 0.28)**	**10.0‡ (*d* = 0.33)**	**7.3‡ (*d* = 0.21)**
NOA, NT	**3.1‡ (*d* = 0.07)**	**2.6‡ (*d* = 0.07)**	**2.5† (*d* = 0.08)**	**1.5† (*d* = 0.07)**	**10.3‡ (*d* = 0.36)**	**7.0‡ (*d* = 0.18)**	**4.6‡ (*d* = 0.12)**	**3.9‡ (*d* = 0.11)**	**5.3† (*d* = 0.20)**	**5.9‡ (*d* = 0.15)**	**5.2‡ (*d* = 0.17)**	**3.7‡ (*d* = 0.10)**
NNC, NNA	1.2 (*d* = 0.03)	0.6 (*d* = 0.01)	0.2 (*d* = 0.02)	−1.8 (*d* = −0.03)	−0.1 (*d* = −0.01)	−0.7 (*d* = 0.00)	0.0 (*d* = 0.00)	−0.8 (*d* = −0.01)	3.3 (*d* = 0.10)	−4.6 (*d* = −0.10)	−1.6 (*d* = −0.03)	−4.9 (*d* = −0.11)
(b) Dutch HMI												
NC, NT	**4.6‡ (*d* = 0.14)**	**4.4‡ (*d* = 0.13)**	**9.2‡ (*d* = 0.28)**	**8.6‡ (*d* = 0.26)**	**13.0‡ (*d* = 0.60)**	**6.6‡ (*d* = 0.20)**	**4.3† (*d* = −0.14)**	**5.6‡ (*d* = 0.19)**	**10.5‡ (*d* = 0.46)**	**10.4‡ (*d* = 0.35)**	**5.2† (*d* = 0.22)**	**7.7‡ (*d* = 0.24)**
NC, NOA	**−1.9 (*d* = −0.03)**	**−3.8† (*d* = −0.11)**	2.2 (*d* = 0.08)	1.0 (*d* = 0.06)	**7.6‡ (*d* = 0.39)**	−1.9 (*d* = −0.01)	−2.3 (*d* = −0.04)	−0.9 (*d* = 0.00)	**10.3‡ (*d* = 0.50)**	−0.2 (*d* = 0.01)	1.3 (*d* = 0.06)	−1.3 (*d* = −0.04)
NOA, NT	**6.5‡ (*d* = 0.17)**	**8.2‡ (*d* = 0.24)**	**7.0‡ (*d* = 0.20)**	**7.6‡ (*d* = 0.20)**	**5.4* (*d* = 0.21)**	**8.5‡ (*d* = 0.21)**	**6.6‡ (*d* = 0.18)**	**6.5‡ (*d* = 0.20)**	0.2 (*d* = −0.04)	**10.6‡ (*d* = 0.34)**	**3.9* (*d* = 0.16)**	**9.0‡ (*d* = 0.28)**
NNC, NNA	1.9 (*d* = 0.07)	−5.9 (*d* = −0.13)	−2.1 (*d* = −0.05)	−2.7 (*d* = −0.05)	2.0 (*d* = −0.12)	−2.3 (*d* = 0.03)	1.5 (*d* = 0.05)	4.2 (*d* = −0.16)	**9.8* (*d* = 0.34)**	−6.2 (*d* = −0.15)	2.4 (*d* = 0.11)	−4.7 (*d* = −0.13)
(c) Flemish read												
NC, NT	**3.9‡ (*d* = 0.14)**	**3.8‡ (*d* = 0.16)**	**1.8* (*d* = 0.11)**	1.3 (*d* = 0.08)	**14.1‡ (*d* = 0.55)**	**7.5‡ (*d* = 0.24)**	**4.9‡ (*d* = −0.17)**	**2.9† (*d* = 0.14)**	**9.1‡ (*d* = 0.43)**	**8.3‡ (*d* = 0.26)**	**9.4‡ (*d* = 0.36)**	**6.3‡ (*d* = 0.20)**
NC, NOA	**7.2‡ (*d* = −0.22)**	**6.5‡ (*d* = −0.23)**	**6.0‡ (*d* = 0.22)**	**3.2† (*d* = 0.12)**	**13.9‡ (*d* = 0.54)**	**10.8‡ (*d* = −0.30)**	**7.5‡ (*d* = −0.23)**	**6.0‡ (*d* = 0.22)**	**17.1‡ (*d* = 0.68)**	**9.5‡ (*d* = 0.29)**	**13.3‡ (*d* = 0.46)**	**7.9‡ (*d* = −0.23)**

Table 7: (continued)

	GC	GT	MNL	MVL	MMS	WV2	WV3	WV3T	NM	CF	CWV2	CX
(a) Dutch read												
NOA,	**−3.3‡**	**−2.7***	**−4.2†**	−1.9	0.2	−3.3	**−2.6***	−3.1	**−8.0†**	−1.2	**−3.9***	−1.6
NT	**(*d* = −0.08)**	**(*d* = −0.07)**	**(*d* = −0.11)**	(*d* = −0.04)	(*d* = 0.00)	(*d* = −0.06)	**(*d* = −0.06)**	(*d* = −0.08)	**(*d* = −0.25)**	(*d* = −0.02)	**(*d* = −0.10)**	(*d* = −0.03)
NNC,	**8.8‡**	**5.7†**	**6.8‡**	3.3	**12.2†**	**10.1†**	**8.8‡**	**5.7***	**11.7‡**	4.6	**12.5‡**	**5.4***
NNA	**(*d* = 0.23)**	**(*d* = −0.16)**	**(*d* = −0.20)**	(*d* = 0.09)	**(*d* = −0.43)**	**(*d* = 0.21)**	**(*d* = 0.23)**	**(*d* = −0.18)**	**(*d* = 0.44)**	(*d* = −0.14)	**(*d* = 0.37)**	**(*d* = −0.14)**
(d) Flemish HMI												
NC,	**6.3†**	−0.2	5.0	**5.9†**	**21.8‡**	**8.6†**	**4.9‡**	**2.9†**	**13.7‡**	**11.1‡**	**15.3‡**	**9.6‡**
NT	**(*d* = 0.18)**	(*d* = 0.00)	(*d* = 0.13)	**(*d* = 0.17)**	**(*d* = 0.86)**	**(*d* = 0.23)**	**(*d* = −0.17)**	**(*d* = 0.14)**	**(*d* = 0.50)**	**(*d* = 0.26)**	**(*d* = 0.45)**	**(*d* = 0.22)**
NC,	4.4	−2.6	3.6	3.4	**22.5‡**	4.5	**7.5‡**	**6.0‡**	**22.2‡**	5.9	**9.4†**	4.3
NOA	(*d* = −0.11)	(*d* = −0.09)	(*d* = 0.07)	(*d* = 0.08)	**(*d* = 0.88)**	(*d* = −0.12)	**(*d* = −0.23)**	**(*d* = 0.22)**	**(*d* = 0.83)**	(*d* = 0.13)	**(*d* = 0.27)**	(*d* = −0.09)
NOA,	1.9	2.4	1.4	2.5	−0.7	4.1	**−2.6***	−3.1	**−8.5†**	**5.2***	**5.9***	5.3
NT	(*d* = −0.07)	(*d* = 0.08)	(*d* = −0.06)	(*d* = −0.09)	(*d* = −0.02)	(*d* = −0.11)	**(*d* = −0.06)**	(*d* = −0.08)	**(*d* = −0.34)**	**(*d* = −0.13)**	**(*d* = −0.18)**	(*d* = −0.13)
NNC,	**6.7†**	**−11.1†**	3.1	0.7	−4.1	6.7	**8.8‡**	**5.7***	**12.2‡**	−7.7	**6.5***	**−2.2***
NNA	**(*d* = 0.19)**	**(*d* = −0.29)**	(*d* = −0.10)	(*d* = 0.04)	(*d* = 0.12)	(*d* = 0.20)	**(*d* = 0.23)**	**(*d* = −0.18)**	**(*d* = 0.59)**	(*d* = −0.15)	**(*d* = 0.29)**	**(*d* = −0.00)**

fewer age performance disparities observed in Dutch non-native speech compared to Flemish non-native speech.

3.2.2 The effect of gender on recognition performance

Table 8 shows the size of the performance disparities between the male and female speaker groups for the different systems and models, split for the five speaker groups for Dutch and Flemish and for read and HMI. Overall, most current state-of-the-art ASR systems and models recognized the female and male speech equally well, except for one case for Dutch, where Whisper-larger-V3 (WV3) exhibited a negligible but significant performance disparity between female and male speakers for native teenagers in HMI speech (d = 0.16, the row of NT of Table 8b).

This picture is in line with the literature which also shows varying results regarding gender performance disparities in different languages and databases. For instance, no gender performance disparities were found for English (Tatman and Kasten 2017) and French (Garnerin et al. 2019), while a slightly better performance on female speech was observed by (Feng et al. 2024; Fuckner et al. 2023; Herygers et al. 2023) for Dutch on the same database, for English (Adda-Decker and Lamel 2005; Goldwater et al. 2008; Koenecke et al. 2020; Raes et al. 2024; Serditova et al. 2025), for Arabic (Sawalha and Shariah 2013), and for French (Adda-Decker and Lamel 2005). Conversely, Tatman (2017) observed a slight worse performance for male speech for English and Garnerin et al. (2019) for French.

3.2.3 The effect of non-native accents on recognition performance

Overall, for Dutch read and HMI speech (top panels of Figure 3), non-native-accented speech (yellow and orange markers) led to higher WERs than native speech (light-blue, gray, and purple markers). For Flemish (bottom panels), most systems and models performed better on native speech than non-native-accented speech.

Table 9 lists the size of the performance disparity between native and non-native children and adults in the tested systems and models for Dutch (Table 9a and b) and Flemish (Table 9c and d) and read and HMI speech. Non-native children were compared with native children. Non-native adults were compared with native teenagers: while native adults are not included in the Jasmin corpus, the WERs reported in Zhang et al. (2023a) showed that Dutch ASR systems performed similarly on native adults' and teenagers' speech. For Dutch read and HMI speech, all systems and models showed significantly worse performance for non-native children and adults. For Flemish read and HMI speech, most systems and models showed significantly worse performance for non-native children and adults. Within each system or model, for the same speaking style and the same speaker group, the performance

Table 8: Performance disparities between male and female speakers in absolute WER differences (%). A positive WER difference means that female speech was recognized better than male speech and vice versa. Bold denotes the performance disparity is significant. Statistical significance levels: *$p < 0.05$; †$p < 0.01$; ‡$p < 0.001$. *d* denotes the effect size.

	GC	GT	MNL	MVL	MMS	WV2	WV3	WV3T	NM	CF	CWV2	CX
(a) Dutch read												
NC	0.1	−0.3	0.1	0.7	−0.5	0.4	1.0	0.5	−2.5	0.1	0.6	0.2
	(*d* = 0.01)	(*d* = −0.01)	(*d* = 0.00)	(*d* = 0.02)	(*d* = 0.00)	(*d* = 0.03)	(*d* = 0.03)	(*d* = 0.02)	(*d* = −0.06)	(*d* = 0.00)	(*d* = 0.02)	(*d* = 0.00)
NT	1.6	1.4	1.4	1.5	4.9	3.2	2.1	2.6	2.8	2.4	2.6	2.6
	(*d* = 0.03)	(*d* = 0.03)	(*d* = 0.04)	(*d* = 0.04)	(*d* = 0.17)	(*d* = 0.07)	(*d* = 0.50)	(*d* = 0.07)	(*d* = 0.09)	(*d* = 0.06)	(*d* = 0.08)	(*d* = 0.07)
NOA	2.9	2.8	2.5	2.5	1.4	5.4	4.0	4.6	6.0	1.4	3.6	1.6
	(*d* = 0.06)	(*d* = 0.07)	(*d* = 0.07)	(*d* = 0.06)	(*d* = 0.06)	(*d* = 0.13)	(*d* = 0.09)	(*d* = 0.12)	(*d* = 0.20)	(*d* = 0.04)	(*d* = 0.12)	(*d* = 0.04)
NNC	4.0	1.9	−0.6	1.1	9.3	3.0	2.5	3.6	3.8	0.8	1.7	0.1
	(*d* = 0.09)	(*d* = 0.06)	(*d* = 0.00)	(*d* = 0.05)	(*d* = 0.34)	(*d* = 0.09)	(*d* = 0.07)	(*d* = 0.11)	(*d* = 0.13)	(*d* = 0.04)	(*d* = 0.07)	(*d* = 0.02)
NNA	3.0	3.9	0.3	−0.4	8.5	−0.2	2.0	2.1	3.6	0.6	−0.1	0.4
	(*d* = 0.08)	(*d* = 0.11)	(*d* = 0.03)	(*d* = 0.01)	(*d* = 0.32)	(*d* = 0.02)	(*d* = 0.07)	(*d* = 0.07)	(*d* = 0.14)	(*d* = 0.03)	(*d* = 0.02)	(*d* = 0.03)
(b) Dutch HMI												
NC	−0.5	0.2	0.2	−3.2	−2.8	0.6	−0.3	0.1	−3.5	−0.3	1.4	−1.5
	(*d* = −0.03)	(*d* = −0.03)	(*d* = −0.05)	(*d* = −0.12)	(*d* = −0.23)	(*d* = −0.04)	(*d* = −0.04)	(*d* = −0.05)	(*d* = −0.24)	(*d* = −0.15)	(*d* = 0.01)	(*d* = −0.01)
NT	3.0	0.6	2.6	2.4	5.0	2.2	**5.2***	4.6	0.3	−0.3	2.2	2.7
	(*d* = 0.08)	(*d* = 0.03)	(*d* = 0.08)	(*d* = 0.06)	(*d* = 0.24)	(*d* = 0.07)	**(*d* = 0.16)**	(*d* = 0.15)	(*d* = 0.07)	(*d* = 0.01)	(*d* = 0.06)	(*d* = 0.09)
NOA	3.2	3.2	3.8	3.0	3.2	3.9	4.2	5.2	4.8	2.5	4.8	3.6
	(*d* = 0.12)	(*d* = 0.12)	(*d* = 0.14)	(*d* = 0.10)	(*d* = 0.19)	(*d* = 0.14)	(*d* = 0.13)	(*d* = 0.19)	(*d* = 0.27)	(*d* = 0.08)	(*d* = 0.18)	(*d* = 0.12)
NNC	2.7	1.2	1.6	1.1	5.5	3.3	1.9	3.8	4.1	3.0	−0.4	1.3
	(*d* = 0.07)	(*d* = 0.01)	(*d* = 0.03)	(*d* = 0.01)	(*d* = 0.54)	(*d* = −0.07)	(*d* = 0.05)	(*d* = 0.16)	(*d* = 0.11)	(*d* = 0.03)	(*d* = −0.04)	(*d* = −0.01)
NNA	4.4	3.6	2.7	2.4	11.7	11.6	5.2	5.4	5.3	4.0	3.2	3.8
	(*d* = 0.12)	(*d* = 0.12)	(*d* = 0.11)	(*d* = 0.08)	(*d* = 0.17)	(*d* = 0.24)	(*d* = 0.14)	(*d* = 0.16)	(*d* = 0.21)	(*d* = 0.05)	(*d* = 0.12)	(*d* = 0.13)
(c) Flemish read												
NC	2.9	2.6	3.6	3.2	4.4	5.5	4.1	3.3	5.6	4.2	3.9	4.2
	(*d* = 0.09)	(*d* = 0.09)	(*d* = 0.13)	(*d* = 0.10)	(*d* = 0.13)	(*d* = 0.13)	(*d* = 0.13)	(*d* = 0.11)	(*d* = 0.19)	(*d* = 0.12)	(*d* = 0.14)	(*d* = 0.12)

Table 8: (continued)

	GC	GT	MNL	MVL	MMS	WV2	WV3	WV3T	NM	CF	CWV2	CX
(a) Dutch read												
NT	2.9	2.0	3.2	2.9	8.0	4.3	4.4	2.8	7.1	3.7	4.2	3.2
	(*d* = 0.07)	(*d* = 0.05)	(*d* = 0.09)	(*d* = 0.07)	(*d* = 0.23)	(*d* = 0.07)	(*d* = 0.10)	(*d* = 0.07)	(*d* = 0.24)	(*d* = 0.08)	(*d* = 0.11)	(*d* = 0.07)
NOA	1.7	1.9	2.4	1.6	−0.3	2.6	3.3	4.6	5.5	4.6	3.3	3.5
	(*d* = 0.04)	(*d* = 0.05)	(*d* = 0.07)	(*d* = 0.04)	(*d* = −0.03)	(*d* = 0.05)	(*d* = 0.07)	(*d* = 0.11)	(*d* = 0.17)	(*d* = 0.09)	(*d* = 0.08)	(*d* = 0.06)
NNC	−1.0	−1.0	−0.4	−1.0	−3.6	2.0	−0.7	0.0	−0.7	−4.7	−2.7	−2.9
	(*d* = −0.04)	(*d* = −0.06)	(*d* = −0.03)	(*d* = −0.03)	(*d* = −0.15)	(*d* = −0.02)	(*d* = −0.05)	(*d* = −0.03)	(*d* = −0.10)	(*d* = −0.14)	(*d* = −0.11)	(*d* = −0.08)
NNA	1.5	1.4	−0.2	0.0	−2.7	−1.3	1.7	−0.3	−1.1	0.5	1.3	−0.9
	(*d* = 0.04)	(*d* = 0.04)	(*d* = −0.01)	(*d* = 0.00)	(*d* = −0.08)	(*d* = 0.00)	(*d* = 0.04)	(*d* = −0.01)	(*d* = −0.04)	(*d* = 0.01)	(*d* = 0.03)	(*d* = −0.02)
(d) Flemish HMI												
NC	2.7	1.7	3.5	3.4	−3.6	5.0	3.1	0.7	−1.1	0.6	−1.9	2.3
	(*d* = 0.09)	(*d* = 0.07)	(*d* = 0.13)	(*d* = 0.10)	(*d* = 0.01)	(*d* = 0.15)	(*d* = 0.13)	(*d* = 0.02)	(*d* = 0.02)	(*d* = −0.01)	(*d* = −0.07)	(*d* = 0.03)
NT	2.9	1.9	3.0	1.7	9.1	5.0	4.4	0.7	7.5	4.8	5.7	4.2
	(*d* = 0.11)	(*d* = 0.08)	(*d* = 0.11)	(*d* = 0.09)	(*d* = 0.40)	(*d* = 0.19)	(*d* = 0.16)	(*d* = 0.11)	(*d* = 0.33)	(*d* = 0.17)	(*d* = 0.15)	(*d* = 0.14)
NOA	2.5	1.8	0.9	1.8	−2.2	4.1	3.1	1.9	1.3	5.0	2.0	3.2
	(*d* = 0.04)	(*d* = 0.01)	(*d* = 0.00)	(*d* = 0.02)	(*d* = −0.15)	(*d* = 0.04)	(*d* = 0.04)	(*d* = 0.02)	(*d* = −0.01)	(*d* = 0.06)	(*d* = −0.03)	(*d* = 0.04)
NNC	−1.1	0.2	−0.4	−1.0	−5.7	−2.0	−2.9	2.0	−7.1	3.8	1.5	1.0
	(*d* = 0.01)	(*d* = 0.03)	(*d* = 0.03)	(*d* = −0.02)	(*d* = −0.06)	(*d* = 0.04)	(*d* = −0.01)	(*d* = 0.06)	(*d* = −0.19)	(*d* = 0.09)	(*d* = 0.11)	(*d* = 0.02)
NNA	0.4	3.5	−1.5	−2.6	−2.1	−1.6	−0.9	2.2	−2.1	−1.6	−3.6	−3.6
	(*d* = 0.05)	(*d* = 0.05)	(*d* = −0.01)	(*d* = −0.02)	(*d* = −0.06)	(*d* = 0.05)	(*d* = 0.05)	(*d* = 0.14)	(*d* = −0.09)	(*d* = 0.12)	(*d* = 0.06)	(*d* = 0.04)

Table 9: Performance disparities between native and non-native accents in absolute WER differences (%). A positive WER difference means that native speech was better recognized than non-native speech and vice versa. NNC is compared to NC and NNA is compared to NT. Bold denotes that the performance disparity is significant. Statistical significance levels: *$p < 0.05$; †$p < 0.01$; ‡$p < 0.001$. *d* denotes the effect size.

	GC	GT	MNL	MVL	MMS	WV2	WV3	WV3T	NM	CF	CWV2	CX
(a) Dutch read												
NNC	**10.8‡**	**11.0‡**	**8.3‡**	**9.6‡**	**25.9‡**	**10.6‡**	**12.3‡**	**17.9‡**	**20.8‡**	**14.4‡**	**8.9‡**	**16.1‡**
	(*d* = 0.23)	**(*d* = 0.27)**	**(*d* = 0.23)**	**(*d* = 0.26)**	**(*d* = 1.50)**	**(*d* = 0.25)**	**(*d* = 0.29)**	**(*d* = 0.47)**	**(*d* = 0.66)**	**(*d* = 0.36)**	**(*d* = 0.28)**	**(*d* = 0.44)**
NNA	**18.7**	**18.4‡**	**19.4‡**	**20.1‡**	**41.4‡**	**24.1‡**	**21.1‡**	**26.8‡**	**33.3‡**	**35.1‡**	**25.7‡**	**32.0‡**
	(*d* = 0.41)	**(*d* = 0.46)**	**(*d* = 0.54)**	**(*d* = 0.53)**	**(*d* = 0.89)**	**(*d* = 0.56)**	**(*d* = 0.50)**	**(*d* = 0.71)**	**(*d* = 1.14)**	**(*d* = 0.89)**	**(*d* = 0.81)**	**(*d* = 0.86)**
(b) Dutch HMI												
NNC	**11.1‡**	**13.1‡**	**4.8***	**6.5‡**	**17.1‡**	**11.4‡**	**13.1‡**	**17.3‡**	**11.4‡**	**13.9‡**	**14.9‡**	**14.8‡**
	(*d* = 0.30)	**(*d* = 0.36)**	**(*d* = 0.15)**	**(*d* = 0.19)**	**(*d* = 0.75)**	**(*d* = 0.32)**	**(*d* = 0.35)**	**(*d* = 0.54)**	**(*d* = 0.47)**	**(*d* = 0.39)**	**(*d* = 0.49)**	**(*d* = −0.42)**
NNA	**13.8‡**	**23.4‡**	**16.1‡**	**17.8‡**	**28.1‡**	**20.3‡**	**15.9‡**	**18.7‡**	**12.1‡**	**30.5‡**	**17.7‡**	**27.2‡**
	(*d* = 0.37)	**(*d* = 0.63)**	**(*d* = 0.47)**	**(*d* = 0.51)**	**(*d* = 1.24)**	**(*d* = 0.49)**	**(*d* = 0.43)**	**(*d* = 0.58)**	**(*d* = 0.59)**	**(*d* = 0.89)**	**(*d* = 0.59)**	**(*d* = 0.78)**
(c) Flemish read												
NNC	**6.8†**	**5.6***	**7.6†**	**6.9†**	**13.7***	8.4	**9.6‡**	**8.1†**	**9.5***	**11.1†**	**9.2***	**9.9†**
	(*d* = 0.14)	**(*d* = 0.10)**	**(*d* = 0.17)**	**(*d* = 0.14)**	**(*d* = 0.31)**	(*d* = 0.08)	**(*d* = 0.22)**	**(*d* = 0.15)**	**(*d* = 0.23)**	**(*d* = −0.20)**	**(*d* = 0.18)**	**(*d* = 0.16)**
NNA	1.9	**3.7***	2.6	**4.9†**	**15.6‡**	5.8	**6.8†**	**4.2***	**6.9***	**14.8‡**	**6.1***	**10.8‡**
	(*d* = 0.05)	**(*d* = 0.10)**	(*d* = 0.07)	**(*d* = 0.12)**	**(*d* = 0.42)**	(*d* = 0.11)	**(*d* = 0.18)**	**(*d* = 0.10)**	**(*d* = 0.22)**	**(*d* = 0.32)**	**(*d* = 0.16)**	**(*d* = 0.22)**
(d) Flemish HMI												
NNC	3.0	**5.8***	2.5	2.0	0.5	4.0	3.8	5.0	−3.8	1.0	2.4	**2.9***
	(*d* = 0.09)	**(*d* = 0.13)**	(*d*=0.09)	(*d* = 0.07)	(*d* = 0.12)	(*d* = 0.12)	(*d* = 0.11)	(*d* = 0.14)	(*d* = −0.07)	(*d* = 0.10)	(*d* = 0.18)	**(*d* = 0.14)**
NNA	2.6	**16.7‡**	4.4	**7.2†**	**26.4‡**	5.9	5.0	3.5	−2.3	**19.8‡**	**11.2‡**	**14.7‡**
	(*d* = 0.07)	**(*d* = 0.41)**	(*d* = 0.12)	**(*d* = 0.20)**	**(*d* = 0.86)**	(*d* = 0.14)	(*d* = 0.12)	(*d* = 0.11)	(*d* = −0.17)	**(*d* = 0.51)**	**(*d* = 0.34)**	**(*d* = 0.36)**

disparities and corresponding effect sizes were consistently larger in Dutch than in Flemish. Furthermore, the largest performance disparity size observed against non-native speakers (41.4 % in Table 9a) exceeds the size related to age (16.1 % in Table 7a) and gender (11.7 % in Table 8b).

Overall, for Dutch, all systems and models exhibited substantial and significant negative effects of non-native accented speech on recognition performance, which is in line with the results reported in Feng et al. (2024) for Dutch on the same database and in Palanica et al. (2019), Roll and Graham (2025), and Wu et al. (2020) for English. For instance, Roll and Graham (2025) found that 10 recent company-developed systems showed large performance differences between native and non-native speakers, with the size of the gap depending on the L1 of the non-native speaker.

3.2.4 The effect of regional accents on recognition performance

Figure 4 shows the WERs, averaged over all age groups, of all ASR systems and models on native speech split for the four accent regions, and reported separately for Dutch and Flemish, for read and HMI speech. Table 10 shows the maximum performance disparities between native regional accents, which is computed by subtracting the lowest WER (region on the right) from the highest WER (region on the

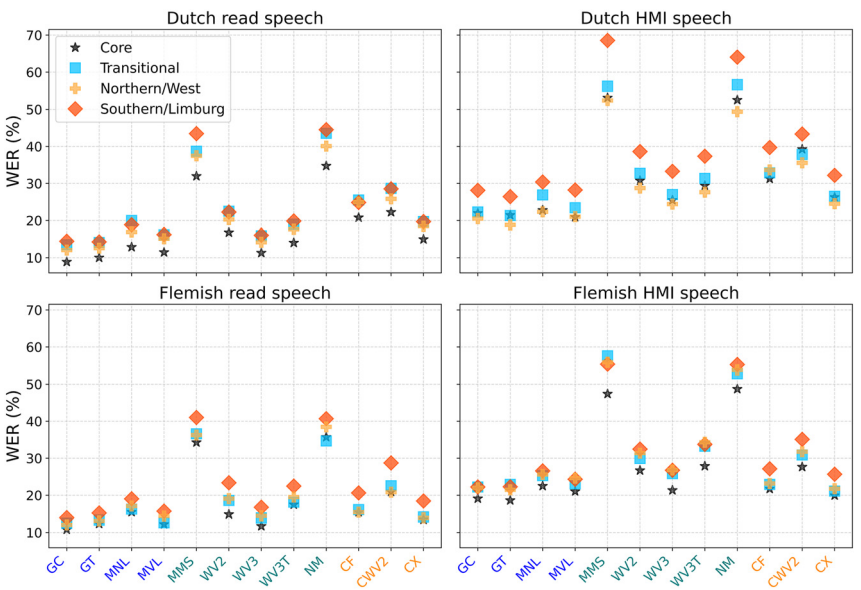

Figure 4: WERs of the ASR systems and models for the native speech averaged over all age groups from the different accent regions in the Netherlands and Flanders.

Table 10: Maximum performance disparities between native regional accents in absolute WER differences (%). WER differences are calculated by subtracting the lowest WER (region on the right) from the highest WER (region on the left) for the four regions in the Netherlands and Flanders within an age group. WER differences are always positive. Bold denotes the performance disparity is significant. Statistical significance levels: *$p < 0.05$; †$p < 0.01$; ‡$p < 0.001$. *d* denotes the effect size.

	GC	GT	MNL	MVL	MMS	WV2	WV3	WV3T	NM	CF	CWV2	CX
(a) Dutch read												
NC	4.0	2.5	2.5	4.0	8.3	7.2	3.7	4.0	7.5	**10.8***	5.7	6.0
	(*d* = 0.07)	(*d* = 0.05)	(*d* = 0.06)	(*d* = 0.09)	(*d* = 0.24)	(*d* = 0.15)	(*d* = 0.07)	(*d* = 0.08)	(*d* = 0.19)	**(*d* = 0.25)**	(*d* = 0.15)	(*d* = 0.15)
	NP, T	NP, T	NP, SP	NP, T	NP, T	NP, SP	NP, T	NP, T	NP, T	**NP, T**	NP, T	NP, T
NT	3.3	2.2	4.8	2.6	5.2	2.7	3.7	2.7	10.2	9.2	5.1	5.4
	(*d* = 0.07)	(*d* = 0.05)	(*d* = 0.13)	(*d* = 0.07)	(*d* = 0.18)	(*d* = 0.06)	(*d* = 0.08)	(*d* = 0.06)	(*d* = 0.31)	(*d* = 0.22)	(*d* = 0.15)	(*d* = 0.14)
	T, NP	T, SP	T, NP	T, NP	T, NP	T, NP	T, SP	T, C	T, NP	T, SP	T, SP	T, SP
NOA	**11.1‡**	**8.4‡**	**9.9‡**	**8.6‡**	**25.5‡**	**16.0‡**	**11.9‡**	**13.1‡**	**19.2‡**	**11.3†**	**14.0‡**	**10.2‡**
	(*d* = 0.24)	**(*d* = 0.20)**	**(*d* = 0.27)**	**(*d* = 0.22)**	**(*d* = 0.88)**	**(*d* = 0.34)**	**(*d* = 0.25)**	**(*d* = 0.31)**	**(*d* = 0.67)**	**(*d* = 0.27)**	**(*d* = 0.42)**	**(*d* = 0.26)**
	SP, NP	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**
(b) Dutch HMI												
NC	3.2	2.4	4.4	2.4	3.9	1.3	0.9	3.3	6.5	7.0	1.5	5.1
	(*d* = 0.09)	(*d* = 0.04)	(*d* = 0.13)	(*d* = 0.08)	(*d* = 0.14)	(*d* = 0.12)	(*d* = 0.06)	(*d* = 0.12)	(*d* = 0.21)	(*d* = 0.18)	(*d* = 0.12)	(*d* = 0.15)
	NP, T	SP, NP	T, SP	NP, SP	SP, T	NP, T	NP, T	NP.T	T, NP	SP, T	NP, T	SP, T
NT	3.5	2.4	3.6	4.1	10.6	4.8	3.4	3.9	9.6	4.3	8.7	1.4
	(*d* = 0.03)	(*d* = 0.03)	(*d* = 0.08)	(*d* = 0.07)	(*d* = 0.21)	(*d* = 0.10)	(*d* = 0.09)	(*d* = 0.09)	(*d* = 0.32)	(*d* = –0.01)	(*d* = 0.26)	(*d* = –0.01)
	SP, T	SP, T	SP, C	SP, C	SP, T	NP, T	NP, T	SP, C	SP, NP	NP, T	SP, T	C, T
NOA	11.5	10.3	**12.0***	**11.3***	**23.2‡**	**15.6***	14.2	**16.0***	**19.9†**	9.4	12.2	10.5
	(*d* = 0.20)	(*d* = 0.19)	**(*d* = 0.25)**	**(*d* = 0.23)**	**(*d* = 0.80)**	**(*d* = 0.35)**	(*d* = 0.26)	**(*d* = 0.35)**	**(*d* = 0.62)**	(*d* = 0.24)	(*d* = 0.30)	(*d* = 0.23)
	SP, NP	SP, NP	**SP, NP**	**SP, NP**	**SP, NP**	**SP, NP**	SP, NP	**SP, NP**	**SP, NP**	SP, T	SP, NP	SP, NP
(c) Flemish read												
NC	3.5	5.2	5.9	4.7	11.0	**14.0***	**8.6***	7.0	8.4	**11.1***	**15.9†**	**9.8***
	(*d* = 0.12)	(*d* = 0.18)	(*d* = 0.21)	(*d* = 0.15)	(*d* = 0.33)	**(*d* = 0.34)**	**(*d* = 0.24)**	(*d* = 0.23)	(*d* = 0.24)	**(*d* = 0.33)**	**(*d* = 0.54)**	**(*d* = 0.29)**
	LP, C	LP, WP	LP, WP	LP, WP	LP, T	**LP, C**	**LP, C**	LP, C	LP, T	**LP, WP**	**LP, WP**	**LP, WP**

Table 10: (continued)

	GC	GT	MNL	MVL	MMS	WV2	WV3	WV3T	NM	CF	CWV2	CX
(a) Dutch read												
NT	4.8	3.2	4.5	3.7	6.8	9.0	4.6	6.2	6.6	4.7	7.1	4.7
	(*d* = 0.11)	(*d* = 0.07)	(*d* = 0.11)	(*d* = 0.08)	(*d* = 0.19)	(*d* = 0.16)	(*d* = 0.10)	(*d* = 0.14)	(*d* = 0.21)	(*d* = 0.10)	(*d* = 0.17)	(*d* = 0.09)
	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, T	LP, C	LP, C	LP, C
NOA	2.5	2.4	3.6	3.8	3.9	3.2	3.3	3.3	1.5	1.5	2.2	1.3
	(*d* = 0.06)	(*d* = 0.05)	(*d* = 0.09)	(*d* = 0.08)	(*d* = 0.12)	(*d* = 0.05)	(*d* = 0.06)	(*d* = 0.08)	(*d* = 0.03)	(*d* = 0.02)	(*d* = 0.04)	(*d* = 0.00)
	WP, LP	WP, C	WP, C	WP, C	T, WP	WP, C	WP, C	WP, T	WP, T	WP, T	T, C	WP, C
(d) Flemish HMI												
NC	16.7	10.5	15.6	11.1	20.9	15.8	12.8	13.0	20.0	13.7	22.3	16.3
	(*d* = 0.29)	(*d* = 0.18)	(*d* = 0.28)	(*d* = 0.16)	(*d* = 0.42)	(*d* = 0.29)	(*d* = 0.24)	(*d* = 0.28)	(*d* = 0.53)	(*d* = 0.19)	(*d* = 0.46)	(*d* = 0.22)
	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C	LP, C
NT	7.8	5.8	4.3	5.5	15.8	8.7	7.2	7.5	8.9	3.5	3.8	1.9
	(*d* = 0.09)	(*d* = 0.03)	(*d* = 0.02)	(*d* = 0.01)	(*d* = 0.42)	(*d* = 0.14)	(*d* = 0.13)	(*d* = 0.15)	(*d* = 0.09)	(*d* = 0.08)	(*d* = 0.18)	(*d* = 0.10)
	WP, LP	WP, LP	WP, C	WP, LP	WP, C	WP, C	WP, C	WP, C	WP, T	T, C	LP, C	LP, C
NOA	2.6	3.4	2.2	3.4	8.8	1.8	4.6	5.5	7.2	3.8	5.2	4.4
	(*d* = 0.04)	(*d* = 0.10)	(*d* = −0.07)	(*d* = −0.03)	(*d* = −0.37)	(*d* = −0.28)	(*d* = −0.21)	(*d* = −0.02)	(*d* = −0.39)	(*d* = −0.16)	(*d* = −0.13)	(*d* = −0.16)
	T, C	LP, C	WP, C	WP, C	T, C	LP, WP	LP, C	WP, C	LP, C	LP, C	T, C	LP, C

left) for the four regions in the Netherlands and Flanders within an age group. WER differences are therefore always positive.

For Dutch, overall, there are no large recognition performance differences between the different regions for children and teenagers, although some of these smaller differences are significant as shown by the performance disparities sizes listed in Table 10. Overall, significant performance disparities between accent regions were observed for all models for the Dutch older adults, where speakers from the southern peripheral region (SP; orange diamonds) were recognized worst (rows NOA of Table 10a and b; top panels of Figure 4). For Flemish, no large regional performance disparities were observed; although five cases of performance disparities were significant with small to medium effect sizes (see the row NC of Table 10c).

Overall, for both Dutch and Flemish, the performance disparities between regional accents are not as frequently observable nor as large as the performance disparities due to non-native accents and age. However, Weilinghoff (2025) reported that most Whisper models achieved lower WERs on Scottish English than on Nigerian English, indicating regional performance disparities, and Feng et al. (2024) also found regional performance disparities in Mandarin.

3.3 The effect of system settings

3.3.1 The effect of data processing pipeline and decoding parameters on recognition performance of Whisper models

In this work, similar to Fuckner et al. (2023), we tested Whisper-large-V2 (WV2) on Jasmin; however, we obtained worse results – our WERs on Dutch and Flemish read and HMI speech ranged from 23.0 % to 37.4 % (row WV2 in Table 4), while Fuckner et al. (2023) reported WERs ranging from 13.3 % to 20.5 % (not shown in the tables of this work).

Through personal communication with the first author of Fuckner et al. (2023), we identified two key differences between their and our experimental setups: (1) the Jasmin data processing pipeline: we segmented the audio into short segments (Section 2.1.1), with one utterance per file, while Fuckner et al. (2023) decoded longer audio files (several minutes). (2) Decoding setups: we set the temperature to 0.0 (beam search), with a beam size of 10. Fuckner et al. (2023) used non-zero temperature values: 0.2, 0.4, 0.6, 0.8, 1.0 (greedy search). To investigate the effect of the data processing pipeline and decoding setups, we created long audio files by concatenating the short segments from the same original audio file according to their original time order using SOX (Chris et al. 2021), resulting in long, concatenated

Table 11: WERs (%) of Whisper-large-V2 (WV2) and Whisper-large-V3 (WV3) on Dutch read and HMI speech and Flemish read and HMI speech using the long, concatenated speech segments. Bold indicates the lowest WER for WV2 or WV3 systems for the same test set. Our setting: temp. = 0.0; beam size = 10.

Model	Decoding parameters	Dutch read	Dutch HMI	Flemish read	Flemish HMI
WV2	Fuckner et al. (2023)	14.5	25.4	10.4	20.6
	Radford et al. (2023)	**14.4**	**24.9**	10.2	**20.3**
	Our setting	14.5	25.0	**9.6**	21.3
WV3	Fuckner et al. (2023)	**19.0**	**24.7**	15.9	**20.5**
	Radford et al. (2023)	19.1	25.3	**15.8**	21.3
	Our setting	123.3	107.5	68.9	67.4

speech segments of 20.0–1,139.8 s. Please note that our long audio files are not identical to those in Fuckner et al. (2023) because we aimed to ensure that the only difference between the current and our previous Whisper experiments (see Section 3.1.1) is the duration. The small duration differences are due to different pre-processing strategies resulting in different parts of the audio files to be kept and removed between the two strategies. Next, we used Whisper-large-V2 (WV2) to decode long audio files with Fuckner et al.'s and our decoding parameters. Additionally, we used the original Whisper paper's decoding setup (Radford et al. 2023) – temperatures of (0.0, 0.2, 0.4, 0.6, 0.8, 1.0). The first block of Table 11 presents the WERs.

To start with the effect of the decoding parameters, this effect is relatively small with similar results for the three settings. However, decoding long audio showed a substantial WER reduction compared to decoding short audio (row WV2 of Table 4 vs. rows WV2 of Table 11). This performance gap is likely due to the strong language modeling capability of the WV2 decoder, which can leverage broader linguistic context in long-form speech. The choice of data processing pipeline should however depend on the application: decoding short, utterance-based audio segments (as in our setup) is suitable for voice assistants and interactive systems, where speech is processed turn by turn. Decoding longer audio segments, as used in Fuckner et al. (2023), is more suitable for tasks involving continuous speech, such as lectures or broadcast transcriptions. Thus, while the long-audio setting yields lower WERs on the Jasmin corpus, the utterance-based setting remains representative and practical for some real-world applications.

We then conducted the same set of experiments using Whisper-large-V3 (WV3) to examine the influence of the data processing pipeline and decoding parameters on the latest Whisper model. As shown in the second block of Table 11, when using both Fuckner et al.'s and Radford et al.'s decoding parameters, the performance

improvement from decoding long audio was smaller than that observed for WV2 (on average across the four test sets, the absolute WER improvement was 4.1 % for WV3 and 12.2 % for WV2; row WV3 of Table 4 vs. rows WV3 with Fuckner et al.'s and Radford et al.'s parameters in Table 11. WERs (%) of Whisper-large-V2 (WV2) and Whisper-large-V3 (WV3) on Dutch read and HMI speech and Flemish read and HMI speech using the long, concatenated speech segments. Our setting: Temp. = 0.0; Beam size = 10.). Notably, when using our decoding parameters, WV3 hallucinated severely, yielding WERs greater than 100 % for Dutch and producing degraded results for Flemish compared with those obtained using Fuckner et al.'s decoding parameters. These results clearly show that Whisper's performance is sensitive to (the combination of) the decoding configuration and the employed data processing (see rows WV3 of Table 11). Previous discussions have noted that certain combinations of deterministic decoding and long audio files can produce runaway hallucinations even for high-quality models (Brown 2026). Our results provide empirical evidence supporting this observation and further indicate that Whisper-large-V3 exhibits hallucinations more frequently than Whisper-large-V2 (GitHub Users 2023, 2024; OpenAI 2023b, c).

Overall, when processing long audio, WV2 performed worse on speech from native older adults, native children, and non-native children and adults compared to native teenagers (Fuckner et al. 2023), which was in line with our findings. However, when comparing female and male speech, Fuckner et al. (2023) found that overall, WV2 performed better on female speech; however, we did not find significant performance disparities between female and male speakers.

3.3.2 Does a new release also mean an improved model?

As indicated in Section 2.2, we started our research in 2024. Comparing the results we obtained in 2025 against those from July 2024, we noticed unexpected performance differences between the various versions of the Microsoft Azure system. Figure 5 presents the WERs of the MNL (Microsoft Dutch) and MVL (Microsoft Flemish) systems obtained in June 2024 (blue circles) and July 2025 (red circles) across the five speaker groups for the four speech types: Dutch read and HMI speech and Flemish read and HMI speech. In summary, the 2025 version of the Dutch system (MNL) performed worse than the June 2024 version for 19 out of the 20 cases, with the exception of Flemish non-native adults' HMI speech. In contrast, both versions of the Flemish system (MVL) (blue and red squares, respectively) showed very similar performance across all cases. These results indicate that, after one year of updates, the performance of the Microsoft Azure system for Dutch (MNL) degraded, whereas the performance of the Flemish system (MVL) remained nearly unchanged.

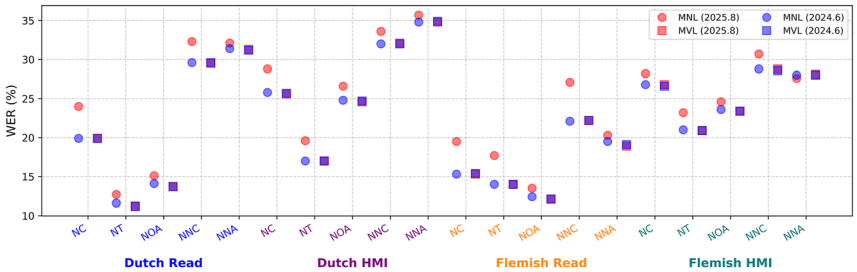

Figure 5: WERs of the MNL and MVL systems for Dutch and Flemish obtained in June 2024 (blue circles) and July 2025 (red circles).

When looking at the performance disparities between different speaker groups, we observed that the performance disparities related to age, gender, non-native and regional accents remained unchanged for the Flemish system. For the Dutch system, performance disparities related to age, gender, and regional accents remained nearly unchanged. However, the performance disparities between non-native adults and native teenagers' speech were reduced for Dutch HMI, Flemish read, and Flemish HMI speech.

4 General discussion and conclusions

Previous research on performance disparities between different speaker groups in Dutch ASR systems has either focused on a single language variety, e.g., Dutch (Feng et al. 2021, 2024) or Flemish (Herygers et al. 2023), or did not separate the two language varieties but treated them as one language (Fuckner et al. 2023). Moreover, previous research examined the performance of several company-developed systems (Fuckner et al. 2023) or trained-from-scratch models (Feng et al. 2021, 2024; Herygers et al. 2023). Our study extended previous research by examining performance disparities in ASR for Dutch and Flemish, separately, evaluating nine company-developed ASR systems from Google, Microsoft, Meta, NVIDIA, and OpenAI, and three custom models trained from scratch on two databases: the Jasmin corpus and the Dutch subset of the DysOne dataset containing speech from a single speaker with severe dysarthria. The three custom models were trained to examine how ASR models trained with transparent and reproducible training data perform on diverse Dutch speech, which aligns with the broader goal of developing fully reproducible (Peng et al. 2023, 2024, 2025) and computationally less-expensive ASR models (Parcollet and Ravanelli 2021).

Among all systems and models, Google Chirp 2 (GC) through API showed the best overall performance across the non-dysarthric speech test sets: Dutch read and human-machine interaction (HMI) speech and Flemish read and HMI speech. GC typically performed best for the native teenagers, native older adults, and non-native adults, across Dutch and Flemish read and HMI speech. For native children, Google Telephony (GT) and the Microsoft Azure system for Flemish (MVL) through API typically performed best. For non-native children, MVL and GT typically performed the best. Of the custom models, the Conformer model using XLSR-53 (CX) features outperformed the Conformer models trained with Whisper and FBank features (CWV2 and CF) on non-dysarthric speech, although the performance disparities related to demographic, language proficiency, and sociolinguistic differences were not reduced compared to the other two models. Importantly, with only ~700 h of training data, the Conformer model using XLSR-53 (CX) features showed similar performance and performance disparity patterns as Whisper-large-V2, showing that with a magnitude less data and computational time needed, similar performance can be obtained for custom models compared to high-performing company-developed models.

Whisper-large-V3 (WV3) performed the best on the dysarthric speech. However, with a WER higher than 75 %, it is still far from usable in practice. Dysarthric speech varies widely in articulation, severity, and etiology, nevertheless, our results for our speaker are in line with extensive results in the literature on severe dysarthric speech recognition. For instance, De Russis and Corno (2019) showed that three company-developed systems yielded 78.2%–89.1 % WERs, and a more recent study showed 49.4%–65.2 % WERs for eight company-developed systems on English severe dysarthric read speech (Alsayegh and Masood 2025). Note that our dysarthric speech also included spontaneous speech, which is worse recognized than read speech (see Section 3.1.3).

This work predominantly focuses on examining performance disparities between different speaker groups classified by demographic factors, assuming that any performance disparities are due to acoustic characteristics shared by the speaker group. However, additional explanations for the found performance disparities exist. There is increasingly more evidence that linguistic, e.g., sentence structures and word categories (Hui et al. 2019; Lopez et al. 2022; Mansfield et al. 2021), and prosodic features, e.g., hesitations and speaking rate (Lopez et al. 2022; Meng et al. 2022), can influence ASR performance differently and systematically across speaker groups. Importantly, different speaker groups may systematically differ in the distribution of the sentence types they produce and are prompted to read (Wan et al. 2024). Therefore, observed ASR performance disparities across demographic groups and speech types may partially reflect differences in linguistic structure and prosodic features in addition to speakers' demographic labels. A systematic analysis of

linguistic and prosodic features will be a promising future work to interpret the sources of performance disparities.

Overall, all systems and models showed highly similar performance disparity patterns. Most performance disparities occurred due to severe speech motor impairment, non-native accents, followed by age, though some performance disparities were found due to differences in regional accents and gender differences. To conclude, our results showed that, for most ASR systems and models, language proficiency differences and severe speech motor impairment had a greater impact on performance disparities than demographic or sociolinguistic factors indicating that acoustic variability due to demographic and sociolinguistic factors is well-represented in "typical speech" training data and consequently is well-modeled in the models. Finally, our analyses of the role of system settings showed that the test data processing pipeline and decoding parameters play an important role in performance and that updates to company-developed systems do not always lead to improvements.

Acknowledgments: The authors thank Dr. Julián Urbano and Dr. Jorge Martinez for their guidance and insightful discussions on statistical significance testing.
Research ethics: The Jasmin and CGN corpora used in this study are publicly available Dutch speech datasets that did not require additional ethical approval. The DysOne dataset was collected by the authors' research group, the Delft Inclusive Speech Communication (DISC) Lab, under approval from the Human Research Ethics Committee (HREC) of Delft University of Technology. Although the author (Y. Zhang) is responsible for collecting the DysOne data, no new data involving human participants were obtained or analyzed for the present study. Accordingly, no additional ethical approval or informed consent was required.
Author contributions: All authors meet the ICMJE criteria for authorship. Y. Zhang: conceptualization, methodology, investigation, validation, writing – original draft; T. De Valck: conceptualization, methodology, investigation, writing – review; O. Scharenborg: conceptualization, supervision, methodology, writing – review and editing. All authors read and approved the final manuscript.
Conflict of interest: The authors have no conflicts of interest to declare.

References

Abouelenin, Abdelrahman, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Dai Qi, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin,

Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang & Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Adda-Decker, Martine & Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Interspeech 2005*, 2205–2208.

Alsayegh, Ali & Tariq Masood. 2025. Zero-shot recognition of dysarthric speech using commercial automatic speech recognition and multimodal large language models. *arXiv preprint arXiv: 2512.17474*.

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers & Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Artificial Analysis. 2024. Speech to Text AI Model & Provider Leaderboard. *Artificial Analysis*. Available at: https://artificialanalysis.ai/speech-to-text.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed & Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33. 12449–12460.

Benjamini, Yoav & Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1). 289–300.

Brown, Emily. 2026. Why Whisper large-v3 Fails on Long Audio. Available at: https://www.technetexperts.com/fix-whisper-large-v3-incomplete-transcription/.

Chang, Xuankai, Takashi Maekaku, Pengcheng Guo, Jing Shi, Yen-Ju Lu, Aswin Shanmugam Subramanian, Tianzi Wang, Shu-Wen Yang, Yu Tsao, Hung-Yi Lee & Shinji Watanabe. 2021. An exploration of self-supervised pretrained representations for end-to-end speech recognition. In *Proceedings of the IEEE automatic speech recognition and understanding workshop (ASRU 2021)*, 228–235. Cartagena, Colombia: IEEE.

Chemudupati, Vamsikrishna, Marzieh Tahaei, Heitor Guimaraes, Arthur Pimentel, Anderson Avila, Mehdi Rezagholizadeh, Boxing Chen & Tiago Falk. 2023. On the transferability of whisper-based representations for" in-the-wild" cross-task downstream speech applications. *arXiv preprint arXiv: 2305.14546*.

Chris, Bagwell, Robs Rullgard Mans & Klauer Ulrich. 2021. SoX – Sound eXchange.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. New York: Routledge.

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed & Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, 2426–2430.

Cucchiarini, Catia, Hugo Van hamme, Olga van Herwijnen & Smits Felix. 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 Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk & Daniel Tapias (eds.), *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). Available at: https://aclanthology.org/L06-1141/.

Davis, S. & P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Transactions on *A*coustics, *S*peech, and *S*ignal *P*rocessing. 28(4). 357–366.

Davitaia, Alexandre. 2025. Applications of machine learning in speech recognition. *International Journal of Artificial Intelligence and Machine Learning* 5(2). 66–69.

De Russis, Luigi & Fulvio Corno. 2019. On the impact of dysarthric speech on contemporary ASR cloud platforms. *Journal of Reliable Intelligent Environments* 5(3). 163–172.

Elhadad, Ahmed, Ibrahim Alrashdi, Abdullah M. Albarrak, Samah Ramadan Ibrahim Elrefaey, Hala Abd Ellatif Elsayed, Farhat Mahmoud Embarak, Zoirov Ulmas & Yousef A. Baker El-Ebiary. 2025. Improved healthcare diagnosis accuracy through the application of deep learning techniques in medical transcription for disease identification. *Alexandria Engineering Journal* 123. 112–123.

Feng, Siyuan, Olya Kudina, Bence Mark Halpern & Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Feng, Siyuan, Bence Mark Halpern, Olya Kudina & Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language* 84. 101567.

Fiscus, Jon. 2015. SCTK: The NIST Scoring Toolkit (sclite). Available at: https://sources.debian.org/data/main/s/sctk/2.4.10-20151007-1312Z%2Bdfsg2-3.1~deb10u1/doc/sclite.htm.

Fuckner, Marcio, Sophie Horsman, Pascal Wiggers & Iskaj Janssen. 2023. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for Dutch speakers. In *2023 International conference on speech technology and human-computer dialogue (SpeD)*, 146–151. IEEE.

Garnerin, Mahault, Solange Rossato & Laurent Besacier. 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*, 3–9.

GitHub Users. 2023. Share your hallucinations here #1873. Available at: https://github.com/openai/whisper/discussions/1873.

GitHub Users. 2024. turbo model release #2363. Available at: https://github.com/openai/whisper/discussions/2363.

Goldwater, Sharon, Dan Jurafsky & Christopher D. Manning. 2008. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates. In Johanna D. Moore, Simone Teufel, James Allan & Sadaoki Furui (eds.), *Proceedings of ACL-08: HLT*, 380–388. Columbus, Ohio: Association for Computational Linguistics. Available at: https://aclanthology.org/P08-1044/.

Google Cloud. 2025. Speech-to-Text: Transcription models. Available at: https://cloud.google.com/speech-to-text/docs/transcription-model.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu & Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of the interspeech*, 5036–5040. International speech communication association (ISCA).

Hernandez, Abner, Paula Andrea Pérez-Toro, Elmar Noeth, Juan Rafael Orozco-Arroyave, Andreas Maier & Seung Hee Yang. 2022. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. In *Interspeech 2022*, 51–55.

Herygers, Aaricia, Vass Verkhodanova, Matt Coler, Odette Scharenborg & Munir Georges. 2023. Bias in Flemish automatic speech recognition. In Christoph Draxler (ed.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, 158–165. Dresden: TUDpress. Available at: https://www.essv.de/pdf/2023_158_165.pdf.

Hui, C. T. Justine, Sahil Jain & Catherine I. Watson. 2019. Effects of sentence structure and word complexity on intelligibility in machine-to-human communications. *Computer Speech & Language* 58. 203–215.

Iancu, Bogdan. 2019. Evaluating Google speech-to-text API's performance for Romanian e-learning resources. *Informatica Economica* 23(1). 17–25.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto & Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys* 55(12). 1–38.

Johnson, Maree, Samuel Lapkin, Vanessa Long, Paula Sanchez, Suominen Hanna, Jim Basilakis & Linda Dawson. 2014. A systematic review of speech recognition technology in health care. *BMC Medical Informatics and Decision Making* 14(1). 94.

Kim, Suyoun, Takaaki Hori & Shinji Watanabe. 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4835–4839.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky & Sharad Goel. 2020. Racial disparities in automated speech recognition. In *Proceedings of the national academy of sciences*, Vol. 117(14), 7684–7689. National Academy of Sciences.

Koenecke, Allison, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann & Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, 1672–1681.

Laaridh, Imed, Corinne Fredouille & Christine Meunier. 2016. Automatic anomaly detection for dysarthria across two speech styles: Read vs spontaneous speech. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 1998–2004.

Latif, Siddique, Junaid Qadir, Adnan Qayyum, Muhammad Usama & Younis Shahzad. 2020. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering* 14. 342–356.

Leeuwen, David A. van, Judith Kessens, E. P. Sanders & Hvd Heuvel. 2009. Results of the N-Best 2008 Dutch speech recognition evaluation. In *Interspeech 2009*, 2571–2574.

Lenth, Russell. 2023. Emmeans: Estimated Marginal Means, aka Least-Squares Means_. *R package version 2.0. 1*.

Lopez, Alianda, Andreas Liesenfeld & Mark Dingemanse. 2022. Evaluation of automatic speech recognition for conversational speech in Dutch, English and German: What goes missing? In Robin Schaefer, Xiaoyu Bai, Manfred Stede & Torsten Zesch (eds.), *Proceedings of the 18th conference on natural language processing (KONVENS 2022)*, 135–143. Potsdam, Germany: KONVENS 2022 Organizers. Available at: https://aclanthology.org/2022.konvens-1.16/.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner & Dominique Makowski. 2021. Performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software* 6(60). 3139.

Luo, Xiao, Le Zhou, Kathleen Adelgais & Zhan Zhang. 2025. Assessing the effectiveness of automatic speech recognition technology in emergency medicine settings: A comparative study of four AI-powered engines. *Journal of Healthcare Informatics Research*. 1–19. https://doi.org/10.1007/s41666-025-00193-w.

Mansfield, Courtney, Sara Ng, Gina-Anne Levow, Richard A. Wright & Mari Ostendorf. 2021. Revisiting parity of human vs. machine conversational speech transcription. In *Interspeech 2021*, 1997–2001.

Meng, Yen, Yi-Hui Chou, Andy T. Liu & Hung-yi Lee. 2022. Don't speak too fast: The impact of data bias on self-supervised speech models. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 3258–3262. IEEE.

Meta, A. I. 2023. facebook/mms-1b-all: Massively Multilingual Speech (MMS). Available at: https://huggingface.co/facebook/mms-1b-all.

Microsoft. 2025. Language and voice support for the Speech service. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/language-support?tabs=stt (Accessed 8 September 2025).

Nelder, John Ashworth & Robert W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society* 135(3). 370–384.

NVIDIA. 2023. stt_nl_fastconformer_hybrid_large_pc. Available at: https://huggingface.co/nvidia/stt_nl_fastconformer_hybrid_large_pc.

NVIDIA. 2024. A New Standard for Speech Recognition and Translation from the NVIDIA NeMo Canary Model. Available at: https://developer.nvidia.com/blog/new-standard-for-speech-\breakrecognition-and-translation-from-the-nvidia-\breaknemo-canary-model/.

Oostdijk, N. 2000. The spoken Dutch corpus. Overview and first evaluation. In *Proceedings of the 2nd international conference on language resources and evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA).

OpenAI. 2023a. Whisper-large-v3. Available at: https://huggingface.co/openai/whisper-large-v3.

OpenAI. 2023c. large-v3 release #1762. Available at: https://github.com/openai/whisper/discussions/1762.

OpenAI. 2023b. turbo model release #2363. Available at: https://github.com/openai/whisper/discussions/2363.

Palanica, Adam, Anirudh Thommandram, Andrew Lee, Michael Li & Yan Fossat. 2019. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. *npj Digital Medicine* 2(1). 55.

Parcollet, Titouan & Mirco Ravanelli. 2021. The energy and carbon footprint of training end-to-end speech recognizers. In *Interspeech 2021*, 4583–4587.

Park, Daniel S., William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk & Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, 2613–2617.

Patel, Tanvina & Odette Scharenborg. 2024. Improving end-to-end models for children's speech recognition. *Applied Sciences* 14(6). 2353.

Peng, Yifan, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-Weon Jung, Soumi Maiti & Shinji Watanabe. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *Automatic speech recognition and understanding workshop (ASRU)*, 1–8. IEEE.

Peng, Yifan, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee-weon Jung & Shinji Watanabe. 2024. OWSM v3.1: Better and faster open whisper-style speech models based on E-Branchformer. In *Interspeech 2024*, 352–356. International speech communication association (ISCA).

Peng, Yifan, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin & Shinji Watanabe. 2025. OWSM v4: Improving open whisper-style speech models via data scaling and cleaning. In *Interspeech 2025*, 2225–2229.

Pratap, Vineel, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve & Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau & Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research* 25(97). 1–52.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey & Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Raes, Rik, Saskia Lensink & Mykola Pechenizkiy. 2024. Everyone deserves their voice to be heard: Analyzing predictive gender bias in ASR models applied to Dutch speech data. *arXiv preprint arXiv:2411.09431*.

Roll, Nathan & Calbert Graham. 2025. Scaling conformation bias in automatic speech recognition. In *Proceedings of the SLaTE 2025*, 133–137.

Russell, Sam O'Connor, Iona Gessinger, Anna Krason, Gabriella Vigliocco & Naomi Harte. 2024. What automatic speech recognition can and cannot do for conversational speech transcription. *Research Methods in Applied Linguistics* 3(3). 100163.

Sapkota, Paban, Abhijit Sinha, Hemant Kumar Kathania & Sudarsana Reddy Kadiri. 2025. Enhancing traditional Kaldi dysarthric speech recognition using SSL-features. In *2025 National Conference on Communications (NCC)* 1–6. https://doi.org/10.1109/NCC63735.2025.10983605.

Sawalha, Majdi & Mohammad Abu Shariah. 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*. Leeds.

Scharenborg, Odette. 2021. Inclusive Speech Technology: Developing Automatic Speech Recognition for Everyone.

Serditova, Dana, Kevin Tang & Jochen Steffens. 2025. Automatic speech recognition biases in Newcastle English: An error analysis. In *Interspeech 2025*, 3204–3208.

Tatman, Rachael. 2017. Gender and dialect bias in YouTube's automatic captions. In Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube & Hanna Wallach (eds.), *Proceedings of the first ACL workshop on ethics in natural language processing*, 53–59. Valencia, Spain: Association for Computational Linguistics.

Tatman, Rachael & Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and YouTube automatic captions. In *Interspeech 2017*, 934–938.

Wan, Yan, Mengyi Sun, Xinchen Kang, Jingting Li, Pengfei Guo, Ming Gao & Su-Jing Wang. 2024. CDSD: Chinese dysarthria speech database. In *Interspeech 2024*, 4109–4113.

Wang, Changhan, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino & Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv: 2101.00390*.

Watanabe, Shinji, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala & Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Interspeech 2018*, 2207–2211. International speech communication association (ISCA).

Weilinghoff, Andreas. 2025. Transcribing diverse voices: Using Whisper for ICE corpora. In *Interspeech 2025*, 3359–3363.

Wienrich, Carolin, Clemens Reitelbach & Astrid Carolus. 2021. The trustworthiness of voice assistants in the context of healthcare investigating the effect of perceived expertise on the trustworthiness of voice assistants, providers, data receivers, and automatic speech recognition. *Frontiers in Computer Science* 3. 685250.

Wu, Yunhan, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R. Doyle, Leigh Clark & Benjamin R. Cowan. 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*, 1–9.

Zhang, Yangyong, Lei Xu, Abner Mendoza, Guangliang Yang, Phakpoom Chinprutthiwong & Guofei Gu. 2019. Life after speech recognition: Fuzzing semantic misinterpretation for voice assistant applications. In *Proceedings of the of the network and distributed system security symposium (NDSS'19)*.

Zhang, Yu, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays & Yonghui Wu. 2023a. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Zhang, Yuanyuan, Aaricia Herygers, Tanvina Patel, Zhengjun Yue & Odette Scharenborg. 2023b. Exploring data augmentation in bias mitigation against non-native-accented speech. In *Automatic speech recognition and understanding workshop (ASRU)*, 1–8. IEEE.

Zhang, Yuanyuan, Matthijs Jasper Valkering, Odette Scharenborg & Zhengjun Yue. 2026. *DysOne: A Dutch and non-native English dysarthric audio-video dataset*. Manuscript in preparation.