Interspeech 2024
1-5 September 2024, Kos, Greece

Challenges of German Speech Recognition: A Study on Multi-ethnolectal Speech Among Adolescents

Martha Schubert[1], *Daniel Duran*[2], *Ingo Siegert*[1]

[1]Mobile Dialog Systems, IIKT, Otto von Guericke University, Magdeburg, Germany
[2]Leibniz-Centre General Linguistics (ZAS), Germany

{martha.schubert;ingo.siegert}@ovgu.de, duran@leibniz-zas.de

Abstract

Despite significant advancements in speech recognition systems, challenges persist in accurately interpreting spontaneous speech from underrepresented groups like non-standard speakers or younger individuals. The difficulty increases when these conditions overlap. To further explore this topic, we employ a dataset featuring spontaneous as well as read speech from young speakers in Germany, including both, speakers from mono-ethnic and multi-ethnic backgrounds. Our study involves a comparative analysis of speech recognition performance, incorporating gender considerations, using three distinct Automatic Speech Recognition (ASR) engines: Whisper (OpenAI), NeMo (NVIDIA), and Wav2Vec2.0 (Meta AI). Furthermore, we conduct a comprehensive error analysis on the automatically generated transcripts, employing part-of-speech (POS) tagging. This allows us to discern the word types that pose the greatest challenge for comprehension by the ASR engines.

Index Terms: speech recognition, adolescent speech, bias, multi-ethnolectal speech

1. Introduction

ASR engines play a crucial role in various applications, from virtual assistants to dictation software, as they have improved rapidly in the last years [1]. However, studies have highlighted disparities in the performance of these systems across different linguistic and cultural groups, as well as across gender [2, 3, 4, 5]. Generally speaking, speaker groups that are not well represented in the training dataset of the ASR are more often misrecognized [6] and there is a significant racial, social and cultural bias in ASR systems as described by Ngueajio and Washington in [7]. It is essential to examine these disparities to ensure equitable access to technology for all users.

Although these things have been observed multiple times on the English language [2, 8], they have also been proven for other languages, e.g., Dutch [9, 6] or French [8]. Our objective is to expand these findings on the German language and to assess the performance of ASR engines in recognizing German speech from accented, non-adult speakers. We aim to investigate whether the results of these other studies also apply to our data set. To our knowledge, our study is the first to systematically investigate the recognition of multi-ethnolectal (German) speech with ASRs, while also utilizing POS-tagging.

For our analysis, we leverage a dataset [10] encompassing spontaneous and read speech recordings alongside their manually created transcripts from adolescents including both multi-ethnolectal and non-ethnolectal German speakers in different experimental settings. We analyze the performance of three distinct ASR engines (Whisper [11], NeMo [12], and Wav2Vec2.0 [13]). Our investigation aims to identify the factors impacting the Word Error Rate (WER) and determining which word types pose the biggest challenge for the ASR engines.

2. Related Work

Due to the great importance of the topic, a lot of research has been conducted on it, examining many different languages and types of speech. As found by Koenecke et al. in [2] there are significant differences in recognizing speech from African American speakers in contrast to white Americans, reporting a WER of $35\,\%$ for speech from people of colour and a WER of $19\,\%$ for white speakers, suggesting a sociolinguistic bias. Tatman and Kasten report in [3] that black speakers are misrecognized more often than white speakers. Similarly, Tatman reports increased difficulty in the recognition of speech from people with a Scottish accent and from female speakers when automatically generating YouTube subtitles [4]. Koenecke et al. also report a gender bias, however stating that female speech is recognized better than male speech, just like Adda-Decker and Lamel in [8] and Fuckner et al. in [6]. In their work, Fuckner et al. report recognition of non-native speech as another challenge for ASR engines, stating that the Whisper as well as the Wav2Vec2 model perform significantly worse on non-native Dutch speech than on native speech. Feng et al. present similar findings in [9] and additionally state that read as well as spontaneous adolescent speech is recognized better than speech from older adults and children when observing native speakers, while there is no big difference between non-native children and adults. Another challenge faced by ASR engines are words rarely appearing or entirely absent in the training corpus. As stated by Wirth and Peinl in [14], $19.82\,\%$ of errors made by ASR engines on German language consist of names, anglicisms or loan words, which all appear at low frequency in training corpora. This poses a problem, since those misrecognized words are often of great importance for specific application contexts, for example when proper names are not understood correctly [15]. Similar observations have been made by Hahn et al. in [16], stating that content words that rarely appear in the training datasets are misrecognized most of the time.

3. Methods

3.1. Dataset

General Information: The dataset contains recordings and manually created transcripts of German speech from students (aged between 13 and 21, mean: 16). The dataset includes 44 speakers, evenly divided with 22 males and 22 females. Furthermore, the dataset is divided into a core group and a comparison group. The first group comprises recordings from 32 multilingual speakers, who, in addition to German, speak languages such as Turkish ($31.25\,\%$), Albanian ($12.5\,\%$), or Bosnian

10.21437/Interspeech.2024-1717

(9.38 %) and attended schools situated in high-immigration, multi-ethnic neighbourhoods of Stuttgart. The selection of these students was determined by ratings provided by external students, who were tasked with discerning whether the recordings originated from individuals from multilingual families. The term *multi-ethnolect* is often used in sociolinguistics for the (colloquial) speaking styles associated with second and third generation speakers from immigrant backgrounds which are used "collectively to express their minority status" [17]. The term *Kiezdeutsch* ("Hood German") is often used for the German multi-ethnolect [18, 19]. The comparison group contains recordings from 12 monolingual German speakers. These speakers are students from a German Realschule (secondary school) sharing a comparable social background, but residing in mono-ethnic settings [10].

The subsets are again divided into three different experimental setups: *Diapix* (D), *Interview* (I) and *Leseliste* (L) (reading list). Diapix consists of recordings in which two students were asked to describe a (custom-made) picture to each other and that way find differences in each other's pictures (cf. [20]). Interview contains recordings of open-ended interviews with the students. In the Leseliste setting, the students were asked to read texts out loud containing 88 disjointed individual sentences intended to cover various phonetic phenomena.

The original audio recordings have a mean length of 698.45 s with a standard deviation of 709.09 s (D:mean=638.9 s, SD=282.2 s, max=2010.6 s, min=43.5 s, I:mean=1896.1 s, SD=519.3 s, max=2735.8 s, min=115.7 s, L:mean=170.3 s, SD=51.2 s, max=338.5 s, min=97.7 s). The audio files have undergone manual pre-processing, having been transcribed orthographically using ELAN [21] and timestamped. We segmented the audio files using the manually generated timestamps on turn-base, resulting in shorter segments with a mean duration of 2.07 s and a standard deviation of 0.93 s. Speech data was recorded in quiet rooms under varying acoustic conditions with Marantz PMD661 MK III handheld digital solid-state recorders (at 44.1 kHz, 16-bit uncompressed PCM format) and Sennheiser MKE 2-P clip-on lavalier microphones for each speaker.

Pre-processing: The manual transcripts as well as those created by the ASR engines underwent various modifications, including the following:

- Capitalizations and punctuation have been removed from both kinds of transcripts to eliminate any differences in style.
- All numerical values in both kinds of transcripts were converted to number words to ensure consistency.
- Spelling errors introduced by human transcribers to clarify the subjects' pronunciation were corrected (e.g. "jetztt" becoming "jetzt"). Additionally, words that can be spelled in different ways have been standardized, such as "oke" in the human transcript becoming "okay".
- Words or phrases that are usually understood to carry no semantic meaning and serve as fillers such as "ähh", "ouh" or "hm" have been removed from transcripts to improve accuracy. Those fillers have also been removed from the ASR transcripts to ensure consistency.
- Colloquial language has been standardized. We categorized 222 colloquial expressions, including expressions like "ner" (short for "einer"), "ne" (short for "eine"), "vonnem" (short for "von einem"), or "nix" (short for "nichts"). However, instead of outright removal, we maintained two versions of the dataset: one with the colloquial expressions intact and one without. During the calculation of the Word Error Rate (WER), if a word labelled as colloquial is identified, our code offers a standard German alternative. If this replacement results in a lower WER, it's retained; otherwise, the colloquial word remains. This approach ensures fairness by not penalizing ASR engines for accurately recognizing colloquial language, while also allowing for improvement in recognition accuracy where possible. A complete list of filler words and colloquial words can be found in the supplementary material (https://github.com/Mobile-Dialog-Systeme/Interspeech2024-Challenges-of-German-Speech-Recognition).

Furthermore, while we deleted single letters if they did not convey a particular meaning, we decided to retain insertions and the fragmented words resulting from disfluencies to ensure accuracy.

Discussion: There are no publicly available datasets of multi-ethnolectal German speech. The General Data Protection Regulation places strict requirements on the handling of personal data, especially when it comes to sensitive data such as that of children. As the data set contains recordings of children speaking, it is of the utmost importance to protect their privacy. The children were also not asked for their consent to use their voice recordings in a public dataset. Therefore, the underlying dataset cannot be made public.

3.2. ASR engines

- **Whisper [11]:** We used OpenAI's medium Whisper Encoder-Decoder model. It is a weakly supervised model that has been trained on 680 000 h audio from the internet, two thirds of which are English and 13 344 h are in German. The training dataset is highly diverse, containing audio from many different situations and environments, as well as different speakers and languages.
- **NeMo [12]:** NVIDIA's NeMo model, we used the Conformer-Transducer-ASR, is a supervised model that uses RNNT/Transducer loss/decoder, making it an autoregressive model [22]. The training dataset contains the Multilingual LibriSpeech Dataset (MLS), VoxPopuli and Mozilla CommonVoice7.0 (MCV7.0).
- **Wav2Vec2.0 [13]:** Wav2Vec2.0 has been trained using self supervised learning, only during fine-tuning, labelled data has been used. In our case MCV8.0, MLS, VoxPopuli and the Multilingual TEDx Dataset have been used to train the facebook/wav2vec2- xls-r-1b [23] model on the German language.

Since all ASR engines have been evaluated on the same dataset (MCV), just with different versions, their performance is easily comparable and reveals that NeMo is the best model and Wav2Vec2.0 is the least favourable (see Table 1).

Table 1: *Comparison of ASR Engines - Parameters, Databases and WER*

	NeMo	Whisper	Wav2Vec2.0
Parameters	120M	769M (medium)	1B
Evaluation Database	MCV7.0	MCV9.0	MCV8.0
WER	4.93%	6.4%	10.95%

3.3. Linguistic Evaluation

3.3.1. POS-Tagging

POS-tagging is employed to identify most error-prone words, utilizing the de_core_news_md model from the SpaCy python

library. Using a trained pipeline, it is decided based on the context which POS-tag is most fitting for each word in a text. An explanation of POS-tags is given in Table 2.

Table 2: *Explanation POS-Tags*

Tag	Explanation	Tag	Explanation
INTJ	interjection	X	other
NUM	numeral	AUX	auxiliary
PART	particle	PRON	pronoun
ADJ	adjective	VERB	verb
NOUN	noun	ADP	adposition
PROPN	proper noun	DET	determiner
ADV	adverb	SCONJ	subordination conjunction

3.3.2. *Word Error Rate*

We calculated the WER using the jiwer python library. The WER in connected speech is defined as the "ratio of the number of errors to the number of words input" [24]. Word errors consist of substitutions (S), insertions (I) and deletions (D) while correctly transcribed words are called hits (H).

$$\mathrm{WER} = \frac{S + D + I}{H + S + D}$$

4. Results

4.1. Word Error Rate

4.1.1. *General Observations*

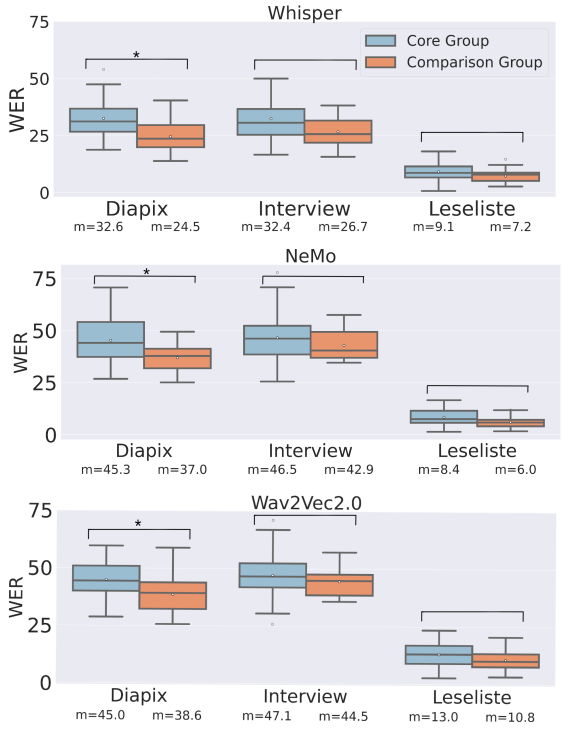

Figure 1: *WER distribution (%) for each model in different experimental settings, statistically significant differences between core and comparison group have been marked using **

Our hypothesis regarding the accuracy of multilingual speakers' speech recognition is supported by our findings. As depicted in Figure 1, the comparison group consistently demonstrates lower WERs across all experimental settings and for all three ASR models compared to the core group (mean difference = 4.58, SD = 2.43). However, t-tests Bonferroni corrected at the local significance level of $0.002\,78\,\%$ reveal that the differences between core and comparison groups are only statistically significant for the Diapix settings for all models (NeMo: t= 3.801, p= 0.0003, Whisper: t= 4.410, p=$3.304{\cdot}10^{-5}$, Wav2Vec2.0: t= 3.535, p= 0.0007). Notably, both core and comparison groups exhibit high WER in the Diapix and Interview settings, indicating challenges in recognizing spontaneous unscripted teenage speech. These results support our assumptions and are consistent with prior research [6, 25, 26].

Furthermore, scripted speech, as in the Leseliste setting, consistently yields lower WERs compared to spontaneous speech, as evidenced in both the Diapix and Interview subsets. The T-tests indicate significant differences between spontaneous speech in the Diapix/Interview and read speech in the Leseliste settings across all models (NeMo: t= -31.937, p= $5.76{\cdot}10^{-84}$, Whisper: t= -22.608, p= $5.831{\cdot}10^{-59}$, Wav2vec2.0: t= -31.865, p= $8.619{\cdot}10^{-84}$) Notably, within spontaneous contexts, the Interview subset exhibits a higher WER compared to the Diapix subset, however the observed differences are not statistically significant. We attribute this slight contrast to the more unrestricted nature of the Interview environment, where participants freely express themselves, resulting in longer and more complex sentences. Conversely, the Diapix setting prompts shorter, more concise responses due to frequent agreement checks between participants, resulting in a lower WER. It is also apparent that Whisper is the best model at recognizing spontaneous speech whereas NeMo is best at read speech. For read speech, this observation matches the WERs provided by the model developers displayed in Table 1, which were calculated on read speech (MCV).

4.1.2. *Gender differences*

Not only are there differences in understanding specific cultural groups, also a gender bias is often prevalent in ASR engines. Therefore, we tested on both the comparison and the core group, if errors occurred more often in male or female speakers.

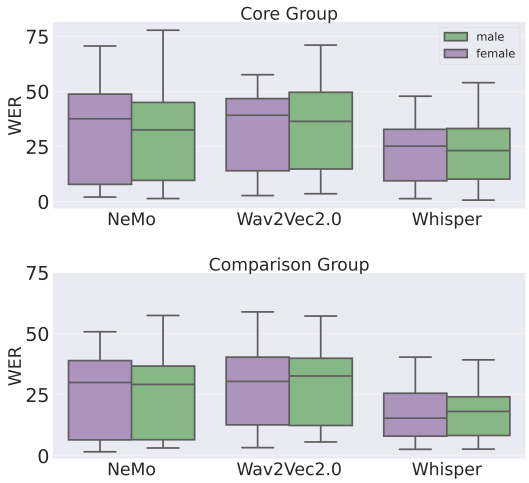

Figure 2: *WER distribution (%) for each model for the core group and the comparison group for female and male speakers*

As depicted in Figure 2, the NeMo model exhibits a slightly

higher Word Error Rate (WER) for female voices compared to male voices, whereas Wav2Vec2.0 and Whisper demonstrates the opposite trend, coinciding with the findings of Fuckner et al. in [6]. However, statistical evaluations using T-tests reveal that the differences between male and female WERs are not statistically significant.

4.2. Colloquial speech

Table 3 shows that colloquial speech is most prevalent in the Diapix setting, followed by the Interview and the Leseliste setting. Interestingly, in the Diapix setting the comparison group shows higher usage of colloquial words than the core group, in contrast to the other settings. This observation can be explained by the fact that many of the speakers in the comparison group, despite German being their mother tongue, have a dialect (i.e. Swabian German) or dialectally influenced German and therefore often saying words like “des” instead of “das”. This dialect appears less in multi-ethnolectal speech (in the core group $53.71\,\%$ of “das” have been pronounced as “des” while in the comparison group it is $79.67\,\%$).

Table 3: *Percentages of colloquial words in total words*

	Diapix	Interview	Leseliste
Core Group	$12.05\,\%$	$9.81\,\%$	$2.13\,\%$
Comparison Group	$15.29\,\%$	$6.92\,\%$	$1.72\,\%$

Among the ASR models assessed in this study, NeMo is best at recognizing colloquial speech, as 748 colloquial words have been recognized. Consequently, NeMo’s transcript aligns more closely with the original speech compared to the transcripts generated by the other two ASR systems. Wav2Vec2.0 follows with 719 recognized colloquial words, while Whisper lags behind with only 565 recognized instances. All models recognized ”des” among their top three colloquial words. Shortenings such as “sag” and “komm” as well as “is” and “komm” were recognized by Whisper and Wav2Vec2.0, while NeMo even recognized ”ne” and ”n”. However, since the overall WER of Whisper is still lower than that of Wav2Vec2.0 and for spontaneous speech also lower than that of NeMo, it becomes apparent that, despite NeMo staying close to what exactly was said, Whisper stands out for its ability to refine and render colloquial speech into standard German. This observation underscores Whisper’s effectiveness in enhancing the clarity of the original speech content.

Table 4: *Percentages of colloquial speech recognized by the ASR engines of total number of colloquial words*

	Whisper	NeMo	Wav2Vec2.0
Core Group	$4.09\,\%$	$4.38\,\%$	$4.75\,\%$
Comparison Group	$3.21\,\%$	$7.41\,\%$	$5.46\,\%$

Additionally, as depicted in Table 4, colloquial speech gets recognized better within the comparison group compared to the core group. This suggests that ASR engines exhibit greater proficiency in deciphering colloquial speech when uttered by mono-lingual, native speakers. Conversely, when colloquial speech is delivered by multi-ethnolectal speakers (core group), recognition rates tend to diminish.

4.3. POS-Tagging

As shown in Figure 3 the words that were incorrectly transcribed are labeled as “X” (other words) and proper nouns, likely due to their absence in the training corpus, aligning with

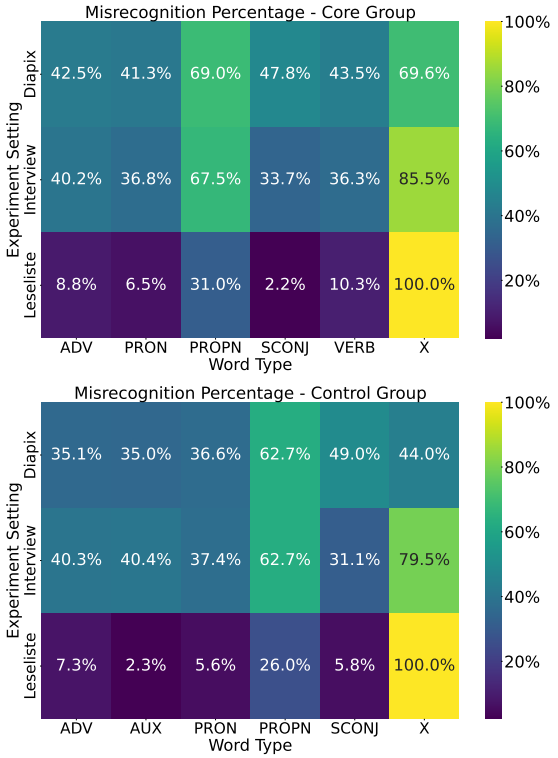

Figure 3: *Heatmap showing most misrecognized word types in core group (top) and cmparison group (bottom), exemplary for NeMo.*

findings by Wirth and Peinl [14] and Ma et al. [15]. Despite the Leseliste setting generally yielding the lowest WER, $100\,\%$ of ’other words’ are transcribed incorrectly, likely because they rarely appear in the texts participants read.

Pronouns additionally pose a difficult word type to recognize by ASR engines. We hypothesize that this is due to their shortness in the German language, consisting only of few letters they are easier to misunderstand. Additionally, they are often differently used by multi-ethnolectal speakers [27], likely increasing the challenge for the ASRs.

5. Conclusion and Discussion

While every effort was made to meticulously prepare the dataset, the scale of the data introduces the potential for minor errors. However, we maintain confidence that these errors are unlikely to significantly diminish the overall reliability of our findings. Our analysis underscores the challenges faced by ASR engines in transcribing spontaneous multi-ethnolectal speech. We identified difficulties ASR engines face when transcribing unknown words. However, our findings indicate no significant difference in ASR performance between male and female voices, suggesting fairness in gender representation. We examined colloquial speech among multi- and mono-ethnic German speakers, noting NeMo’s accuracy and Whisper’s proficiency in converting it to standard German. Moving forward, refining ASR algorithms is essential to improve inclusivity in speech recognition technology, particularly in practical domains such as educational technology and healthcare or accessibility systems, where equitable representation and accuracy are vital.

6. Acknowledgements

The collection of the dataset of multiethnolectal (Stuttgart) German has been supported by the German Research Foundation (DFG, grant Au 72/27-1, PI: Peter Auer, at the University of Freiburg). We thank the schools in Stuttgart for their kind support and our speakers for their participation.

7. References

[1] A. S. Dhanjal and W. Singh, "A comprehensive survey on automatic speech recognition using neural networks," *Multimedia Tools and Applications*, Aug. 2023.

[2] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, p. 7684–7689, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1073/pnas.1915768117

[3] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *Proc. Interspeech 2017*, 2017, pp. 934–938.

[4] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, and H. Wallach, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 53–59. [Online]. Available: https://aclanthology.org/W17-1606

[5] I. Siegert, Y. Sinha, O. Jokisch, and A. Wendemuth, *Recognition Performance of Selected Speech Recognition APIs – A Longitudinal Study*. Cham: Springer International Publishing, 2020, pp. 520–529.

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

[7] M. K. Ngueajio and G. Washington, "Hey asr system! why aren't you more inclusive?" in *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, J. Y. C. Chen, G. Fragomeni, H. Degen, and S. Ntoa, Eds. Cham: Springer Nature Switzerland, 2022, pp. 421–440.

[8] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *Proc. Interspeech 2005*, 2005, pp. 2205–2208.

[9] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," 2021. [Online]. Available: https://arxiv.org/abs/2103.15122

[10] P. Auer and D. Duran, "Coronalisation in the German multiethnolect: Evidence for regional differentiation?" in *The Continuity of Linguistic Change: Selected papers in honour of Juan Andrés Villena-Ponsoda*, ser. Studies in Language Variation, M. Vida-Castro and A. M. Ávila Muñoz, Eds. Amsterdam: John Benjamins Publishing Company, 2024, no. 31, pp. 79–99. [Online]. Available: https://doi.org/10.1075/silv.31.04aue

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[14] J. Wirth and R. Peinl, "Automatic speech recognition in german: A detailed error analysis," 08 2022, pp. 1–8.

[15] X. Ma, X. Wang, and D. Wang, "Low-frequency word enhancement with similar pairs in speech recognition," in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2015, pp. 343–347.

[16] S. Hahn, A. Sethy, H.-K. J. Kuo, and B. Ramabhadran, "A study of unsupervised clustering techniques for language modeling," in *Interspeech*, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:29447280

[17] M. Clyne, "Lingua franca and ethnolects in Europe and beyond," *Sociolinguistica - International Yearbook of European Sociolinguistics*, vol. 14, no. 1, pp. 83–89, Dec. 2000.

[18] H. Wiese, *Kiezdeutsch: ein neuer Dialekt entsteht*, ser. Beck'sche Reihe. München: Beck, 2012, no. 6034.

[19] S. Jannedy and M. Weirich, "Some aspects on individual speaking style features in Hood German," in *Proc. Speech Prosody 2014*, 2014, pp. 843–847.

[20] R. Baker and V. Hazan, "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior Research Methods*, vol. 43, no. 3, pp. 761–770, 2011.

[21] H. Brugman and A. Russel, "Annotating multi-media/multimodal resources with ELAN," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, Eds. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf

[22] NVIDIA Corporation, "Citedrive brings reference management to overleaf," 2024, https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/asr/models.html [Accessed: Feb 27, 2024].

[23] J. Grosman, "Fine-tuned XLS-R 1B model for speech recognition in German," https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-german, 2022.

[24] A. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," 10 2004.

[25] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (asr) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, no. 9, p. 4419, Apr. 2022. [Online]. Available: http://dx.doi.org/10.3390/app12094419

[26] Silber-Varod, Siegert, Jokisch, Sinha, and Geri, "A cross-language study of selected speech recognition systems," *The Online Journal of Applied Knowledge Management: OJAKM*, vol. 9, pp. 1–15, 2021. [Online]. Available: https://doi.org/10.36965/OJAKM.2021.9(1)1-15

[27] V. Siegel, *Multiethnolektale Syntax: Artikel, Präpositionen und Pronomen in der Jugendsprache*, ser. Oralingua. Heidelberg: Universitätsverlag Winter, 2018, no. 16.