BIAS IN FLEMISH AUTOMATIC SPEECH RECOGNITION

Aaricia Herygers[1], *Vass Verkhodanova*[2], *Matt Coler*[2], *Odette Scharenborg*[3], *Munir Georges*[1,4]
[1]*AImotion Bavaria, Technische Hochschule Ingolstadt,* [2]*Campus Fryslân, University of Groningen,* [3]*Multimedia Computing Group, Delft University of Technology,* [4]*Intel Labs Germany*
aaricia.herygers@thi.de

Abstract: Research has shown that automatic speech recognition (ASR) systems exhibit biases against different speaker groups, e.g., based on age or gender. This paper presents an investigation into bias in recent Flemish ASR. Seeing as Belgian Dutch, which is also known as Flemish, is often not included in Dutch ASR systems, a state-of-the-art ASR system for Dutch is trained using the Netherlandic Dutch data from the Spoken Dutch Corpus. Using the Flemish data from the JASMIN-CGN corpus, word error rates for various regional variants of Flemish are then compared. In addition, the most misrecognized phonemes are compared across speaker groups. The evaluation confirms a bias against speakers from West Flanders and Limburg, as well as against children, male speakers, and non-native speakers.

1 Introduction

The omnipresence of the Internet of Things came with a rise in voice-automated devices such as smart speakers. However, these devices do not perform equally well for everyone trying to use speech to control the room temperature or create a shopping list. The accuracy with which the voice commands are recognized by the automatic speech recognition (ASR) system is dependent on various factors. For example, ASR systems experience difficulties in distinguishing commands when multiple people are speaking [1], when the environment is noisy [2, 3], or when the speech contains multiple languages [4]. Another influence on the accuracy of ASR are the speakers themselves [5].

Previous research, often focusing on the English language [e.g., 6, 7, 8], has shown performance differences in ASR due to sociolinguistic factors such as age and gender. The lower recognition performance for a speaker group compared to another speaker group is referred to as 'bias' [9]. As these systems are used worldwide, finding biases in different languages is of high importance. Therefore, a recent study quantified bias between various speaker groups in a Dutch state-of-the-art ASR system [9]. It demonstrated that an ASR trained on Netherlandic (NL) Dutch performed poorly on Flemish, Dutch spoken in Belgium, while Van Dyck et al. [10] found low word error rates (WERs) for Flemish in a Flemish-trained model, namely approximately 10%. Since most (commercial) ASR systems do not include Flemish [11], and Flemish speakers thus have to use NL Dutch ASR systems, and since Flemish consists of various regional varieties that differ in intelligibility, [e.g., 12], this paper seeks to answer the question: To what extent does a Dutch hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) ASR system exhibit bias against speakers from different Flemish regions, and what is the influence of age, gender, and (non-) nativeness on bias and recognition performance? This question is further investigated through an analysis of the misrecognized phonemes for all speaker groups.

In Section 2 we describe the method. The results from the bias analysis are in Section 3. The results are then discussed in Section 4, followed by a conclusion.

2 Method

The training and testing corpora are described in Section 2.1 and the ASR model is outlined in 2.2. The evaluation is provided in Section 2.3.

2.1 Corpora

The Spoken Dutch Corpus (CGN) [13] was used for training. The CGN contains speech from various scenarios including spontaneous face-to-face conversations, news reports, and telephone conversations. The training data consisted of approximately 500h of NL Dutch.

To evaluate the model, Flemish data from 207 speakers from the JASMIN-CGN corpus [14] was used. The corpus is an extension of the CGN and contains read speech and human-computer interaction (HMI) speech data of children and youngsters, older adults as well as speakers whose first language is not Dutch. The inclusion of this speaking style in the corpus is interesting, given that many applications of ASR, such as smart kitchen devices [15], are designed for (semi-) spontaneous speech rather than read speech.

The test data consisted of speech from four Flemish regions. Table 1 shows the regions and the number of male and female speakers (please note that only binary genders are indicated in the metadata of JASMIN), as well as the various age groups and amount of data per group.

Regional variety	Speakers (female, male)	Group (ages)	Speakers (female, male)	Data
West Flemish (peripheral region)	40 (22, 18)	Native Children (7–11)	43 (23, 20)	6h 10m
East Flemish (transitional region)	38 (19, 19)	Native Youngsters (12–16)	44 (22, 22)	6h 10m
Brabantian (core region)	62 (32, 30)	Native Older Adults (65+)	30 (19, 11)	5h 5m
Limburgish (peripheral region)	34 (18, 16)	Non-native Children (7–16)	52 (25, 27)	6h 10m
		Non-native Adults (18–60)	30 (19, 11)	6h 10m

Table 1 – Number of speakers per region and age group.

For adult non-native speakers, their Dutch language proficiency in the form of their level in the Common European Framework of Reference for Languages (CEFR) was provided. A1: 9 speakers (6 female, 3 male). A2: 9 speakers (5 female, 4 male). B1: 11 (8 female, 3 male).

2.2 ASR System

The state-of-the-art ASR system was a hybrid TDNN-BLSTM DNN-HMM system [16] from Feng et al. [9] which was trained using Kaldi [17]. The TDNN-BLSTM model consisted of three 1024-dimensional TDNN layers and three sets of bi-directional, 1024-dimensional LSTM layers. The lattice-free maximum mutual information criterion [18] is used to train the model, alongside various data augmentation methods to increase the amount of training data: noise [19], speed perturbation [20], and reverberation [21]. High-resolution mel-frequency cepstral coefficients of 40 dimensions are used as input features to the acoustic model, which is trained for 4 epochs. A pre-trained GMM-HMM elicits context-dependent phone alignments through forced alignment, which are used to train the acoustic model. The system utilizes an RNNLM [22] with three TDNN and two LSTM layers. N-best results are generated by a tri-gram language model and rescored by the RNNLM, which are both trained on CGN transcriptions.

2.3 Evaluation

The model was evaluated using samples of read and HMI speech by calculating WERs for the different speaker groups, based on weighted averages. Biases were estimated by investigating

the differences in WERs across speaker groups. Phoneme error rates were obtained from phonetic transcriptions using the Phonemizer [23] with an eSpeak NG[1] backend, scored with the SCLITE Scoring Package[2]. This allowed for a closer examination of which phonemes were susceptible to misrecognition and could cause biases.

3 Results

Overall WERs of 41.97% and 47.90% were obtained for read and HMI speech, respectively. The results from the bias analysis based on different factors are provided in the following sections: region in Section 3.1, age and (non-) nativeness in 3.2, and gender in Section 3.3. Moreover, an analysis of the most misrecognized phonemes is provided in Section 3.4.

Please note that the averages in the figures below are weighted, meaning that the ratio of female to male speakers and their respective WERs impacted the average WERs.

3.1 Regional Biases

Figure 1 shows that, for read speech, speakers from Brabant were best recognized and those from West Flanders and Limburg the worst. A quantification of the bias shows that the model performed 22.4% worse for West Flemish and 20.9% for Limburgish speakers compared to Brabantian speakers.

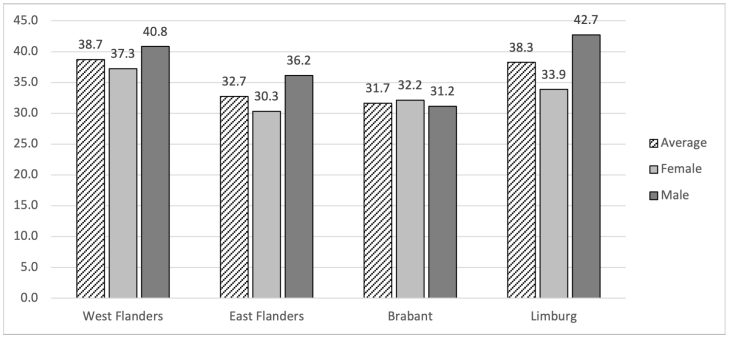

Figure 1 – WERs (%) for read speech across region and gender.

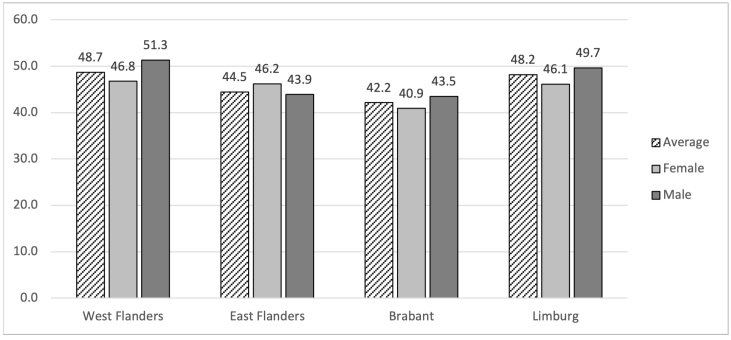

Figure 2 – WERs (%) for HMI speech across region and gender.

From Figure 2 it is clear that speech from HMI was recognized more poorly than read speech, with WERs approximately 10 percentage points higher than in Figure 1. This discrepancy can be found across the different regions. Still, overall, the model performed best for

[1] https://github.com/espeak-ng/espeak-ng

[2] https://github.com/usnistgov/SCTK

Brabantian speech and worst for West Flemish and Limburgish. A comparison of the averages per region found a 15.5% bias against West Flemish speakers and 14.3% against Limburgish speakers compared to Brabantian speakers.

3.2 Age and Non-Nativeness Biases

Figure 3 indicates that native children and youngsters were recognized approximately 65% and 37% more poorly than native older adults, respectively. A comparison of the results for read speech by non-native children to those of native children showed a bias of approximately 27%.

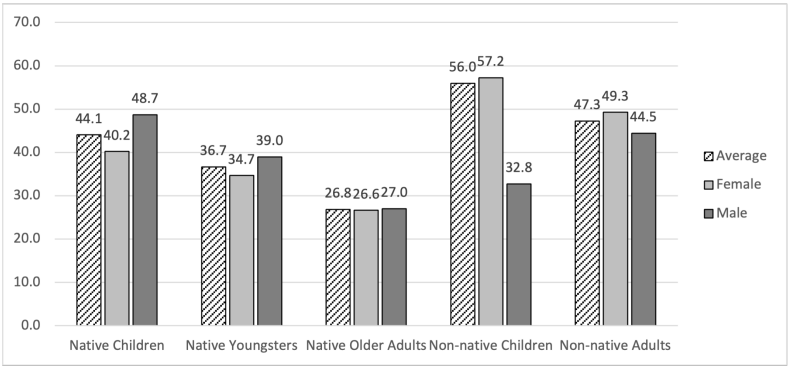

Figure 3 – WERs (%) for read speech across age, gender, and (non-) nativeness.

For HMI, the model also performed best on speech by native older adults, as shown in Figure 4. This resulted in a bias against children of 37.4%. For non-native children and adults, a difference of 7.7% was found. However, a comparison between HMI speech by non-native children and native children showed a smaller bias than for read speech, around 2%.

Table 2 splits the recognition results of non-native speakers up according to their CEFR level and shows that the best WERs were obtained for speakers in the A2 level.

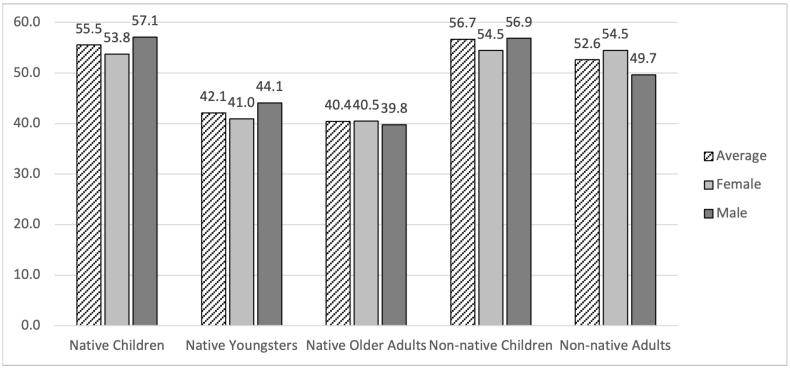

Figure 4 – WERs (%) for HMI speech across age, gender, and (non-) nativeness.

CEFR	Read			HMI		
	Female	Male	Average	Female	Male	Average
A1	52.7	50.2	51.8	60.0	74.1	62.1
A2	40.1	46.4	42.9	42.8	46.5	43.8
B1	52.0	46.4	50.3	59.8	50.3	55.0

Table 2 – WERs (%) for read and HMI speech across language learning levels.

3.3 Gender Bias

Similar to Feng et al. [9], in Figures 1 through 4 we found that across all age groups, male speakers were misrecognized more often (43.54%) than female speakers (40.83%), resulting in a bias against male speakers of 6.6%. However, a smaller bias of 0.7% was found in HMI, with WERs of 48.02% and 47.69% for male and female speakers, resp. Additionally, non-native girls were recognized approx. 75% better than non-native boys in read speech. Smaller differences were found for non-native female and male speech, and for all non-native HMI speech.

3.4 Phoneme Error Analysis

Finally, the results in Table 3 demonstrate that there was no clear difference in the recognition of certain phonemes across regions, genders, age groups, or native versus non-native speakers. Across all regions, /eː/ has the most misrecognitions, followed by /eː/, /ɣ/, /ə/, /ɛ/, and /h/.

Region	Phonemes	Age and Non-Nativeness	Phonemes	Gender	Phonemes
West Flanders	/eː/, /ɛ/, /ə/, /t/, /ɑ/	Native Children	/eː/, /ɣ/, /h/, /ɛ/, /ə/	Female	/eː/, /ɛ/, /ɣ/, /ə/, /t/
East Flanders	/eː/, /ə/, /ɛ/, /h/, /ɣ/	Native Youngsters	/ə/, /eː/, /ɛ/, /t/, /n/	Male	/eː/, /ɣ/, /ə/, /ɛ/, /h/
Brabant	/eː/, /ə/, /ɛ/, /t/, /n/	Native Older Adults	/eː/, /t/, /ɛ/, /ə/, /ɑ/		
Limburg	/eː/, /ɣ/, /ə/, /t/, /ɛ/	Non-native Children	/eː/, /h/, /ɣ/, /ə/, /ɛ/		
		Non-native Adults	/eː/, /ɣ/, /ə/, /ɛ/, /n/		

Table 3 – Misrecognized phonemes across speaker groups

4 Discussion and Conclusion

This paper presented an experiment on bias due to regional language varieties, age, non-nativeness, and gender in Flemish automatic speech recognition using a Dutch-trained DNN-HMM model. By investigating WERs across different groups, we showed that a state-of-the-art Dutch ASR system has biases against West Flemish, Limburgish, young, male, and non-native speakers.

The results from Figures 1 and 2 are in line with studies into human intelligibility that demonstrated that West Flemish and Limburgish speech were perceived as less intelligible than other Flemish and NL Dutch varieties [12, 24, 25]. Both low-performance regions are considered peripheral regions, whereas the high-performing Brabant is considered a central region. The results are in accordance with Impe et al.'s statement that "dialectal language use seems to have preserved quite a strong position" in the peripheral regions [26, p. 104].

While we expected that native youngsters would have the lowest WERs as was found in the NL Dutch data [9], our results showed that seniors were recognized better. The obtained WERs for Flemish older adults were relatively similar to those found by Feng et al. [9], but the error rates for youngsters and children were considerably higher. This might be explained by the increasing linguistic distance between Belgian and NL Dutch [27], meaning that speech by Flemish seniors is closer to speech by Dutch seniors than Flemish minors to Dutch minors.

Perhaps somewhat surprisingly, recognition results were higher for non-native speakers with level A2 than speakers with level B1. One possible explanation may be that the proficiency is typically measured across several factors: speaking, writing, listening, and reading. It is possible that the higher proficiency level of the B1 speakers is not (that much) reflected in their speech production, but rather in their writing, listening, and reading skills. However, the small sample size for each group may impact the generalizability of these results. Additionally, these findings differ from those found by Feng et al. [9] for NL Dutch so more research is warranted.

Gender bias seems to be highly susceptible to the training data, as previous studies have found varying results. For instance Garnerin et al. [28] found that an imbalanced corpus performed better for male speakers, whereas Adda-Decker and Lamel [29] obtained better recognition for female speakers in a balanced corpus. Our findings are thus surprising as the CGN contains more male than female speakers.

Our analysis of the misrecognized phonemes gave different results from the misrecognitions for Flanders found in Feng et al. [9]. They found a high rate of misrecognitions for /y/, /œy/, and /ɑu/, whereas our most misrecognized phonemes were /eː/, /ɣ/, /ə/, /ɛ/, and /h/, which we found in almost all groups. This might be due to the use of another automatic phonetic transcription and phoneme error calculation. However, like Feng et al. [9], we found that /ə/, /h/, and /ɛ/ were misrecognized in many groups. This suggests that these phonemes are difficult for the model to recognize regardless of speaker group or country.

All in all, it is clear that significant steps are needed to reduce the WERs across all 'non-norm speaker groups', as they are considerably higher than those found by Van Dyck et al. [10] in a Flemish-trained model tested on 'norm speech'.

Further research could thus include a larger dataset or more recent architectures as well as bias mitigation techniques. Furthermore, this study could be expanded on by researching biases in a DNN-HMM model trained on Flemish speech or a combination of NL Dutch and Flemish.

References

[1] QIAN, Y.-M., C. WENG, X.-K. CHANG, S. WANG, and D. YU: *Past review, current progress, and challenges ahead on the cocktail party problem*. *Frontiers of Information Technology & Electronic Engineering*, 19(1), pp. 40–63, 2018. doi:https://doi.org/10.1631/FITEE.1700814.

[2] HAMIDI, M., H. SATORI, O. ZEALOUK, and K. SATORI: *Amazigh digits through interactive speech recognition system in noisy environment*. *International Journal of Speech Technology*, 23(1), pp. 101–109, 2020. doi:10.1007/s10772-019-09661-2. URL `https://doi.org/10.1007/s10772-019-09661-2`.

[3] KATHANIA, H. K., S. R. KADIRI, P. ALKU, and M. KURIMO: *Using data augmentation and time-scale modification to improve asr of children's speech in noisy environments*. *Applied Sciences*, 11(18), 2021. doi:10.3390/app11188420. URL `https://www.mdpi.com/2076-3417/11/18/8420`.

[4] SREERAM, G. and R. SINHA: *Exploration of end-to-end framework for code-switching speech recognition task: Challenges and enhancements*. *IEEE Access*, 8, pp. 68146–68157, 2020. doi:10.1109/ACCESS.2020.2986255.

[5] BENZEGHIBA, M., R. DE MORI, O. DEROO, S. DUPONT, T. ERBES, D. JOUVET, L. FISSORE, P. LAFACE, A. MERTINS, C. RIS, R. ROSE, V. TYAGI, and C. WELLEKENS: *Automatic speech recognition and speech variability: A review*. *Speech Communication*, 49(10–11), pp. 763–786, 2007. doi:10.1016/j.specom.2007.02.006. URL `https://doi.org/10.1016/j.specom.2007.02.006`.

[6] KATHANIA, H. K., S. REDDY KADIRI, P. ALKU, and M. KURIMO: *Study of formant modification for children ASR*. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7429–7433. 2020. doi:10.1109/ICASSP40776.2020.9053334.

[7] TATMAN, R. and C. KASTEN: *Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube automatic captions*. In *Proceedings of Interspeech 2017*, pp. 934–938. 2017. doi:10.21437/Interspeech.2017-1746.

[8] VIPPERLA, R., S. RENALS, and J. FRANKEL: *Ageing voices: The effect of changes in voice parameters on ASR performance*. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, pp. 1–10, 2010. doi:https://doi.org/10.1155/2010/525783.

[9] FENG, S., O. KUDINA, B. M. HALPERN, and O. SCHARENBORG: *Quantifying bias in automatic speech recognition*. *arXiv preprint arXiv:2103.15122*, 2021. URL https://doi.org/10.48550/arXiv.2103.15122.

[10] VAN DYCK, B., B. BABAALI, and D. VAN COMPERNOLLE: *A hybrid ASR system for Southern Dutch*. *Computational Linguistics in the Netherlands Journal*, 11, pp. 27–34, 2021. URL https://clinjournal.org/clinj/article/view/119.

[11] SUMMA LINGUAE TECHNOLOGIES: *Language support in voice assistants compared*. https://summalinguae.com/language-technology/language-support-voice-assistants-compared/, 2021. Accessed: 2022-08-09.

[12] VAN BEZOOIJEN, R. and R. VAN DEN BERG: *Taalvariëteiten in Nederland en Vlaanderen: hoe staat het met hun verstaanbaarheid? Taal en Tongval*, 51, pp. 15–33, 1999. URL https://hdl.handle.net/2066/132061.

[13] OOSTDIJK, N.: *The Spoken Dutch Corpus. Overview and first evaluation*. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA), Athens, Greece, 2000. URL http://www.lrec-conf.org/proceedings/lrec2000/pdf/110.pdf.

[14] CUCCHIARINI, C., H. VAN HAMME, O. VAN HERWIJNEN, and F. SMITS: *JASMIN-CGN: Extension of the Spoken Dutch Corpus with speech of elderly people, children and non-natives in the human-machine interaction modality*. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italy, 2006. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/254_pdf.pdf.

[15] KENDRICK, C., M. FROHNMAIER, and M. GEORGES: *Audio-visual recipe guidance for smart kitchen devices*. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pp. 257–261. Association for Computational Linguistics, Trento, Italy, 2021. URL https://aclanthology.org/2021.icnlsp-1.30.

[16] DAHL, G. E., D. YU, L. DENG, and A. ACERO: *Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition*. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), pp. 30–42, 2012. doi:10.1109/TASL.2011.2134090.

[17] POVEY, D., A. GHOSHAL, G. BOULIANNE, L. BURGET, O. GLEMBEK, N. GOEL, M. HANNEMANN, P. MOTLICEK, Y. QIAN, P. SCHWARZ, Y. QIAN, P. SCHWARZ, J. SILOVSKY, G. STEMMER, and K. VESELY: *The Kaldi speech recognition toolkit*. In *Proceedings of the 2011 IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011. URL https://www.danielpovey.com/files/2011_asru_kaldi.pdf.

[18] POVEY, D., V. PEDDINTI, D. GALVEZ, P. GHAHREMANI, V. MANOHAR, X. NA, Y. WANG, and S. KHUDANPUR: *Purely sequence-trained neural networks for ASR based on lattice-free MMI.* In *Proceedings of Interspeech 2016*, pp. 2751–2755. 2016. doi:10.21437/Interspeech.2016-595.

[19] SNYDER, D., G. CHEN, and D. POVEY: *Musan: A music, speech, and noise corpus*. *arXiv preprint arXiv:1510.08484*, 2015.

[20] KO, T., V. PEDDINTI, D. POVEY, and S. KHUDANPUR: *Audio augmentation for speech recognition*. In *Proceedings of Interspeech 2015*. 2015. URL `https://www.isca-speech.org/archive_v0/interspeech_2015/i15_3586.html`.

[21] KO, T., V. PEDDINTI, D. POVEY, M. L. SELTZER, and S. KHUDANPUR: *A study on data augmentation of reverberant speech for robust speech recognition*. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224. IEEE, 2017. doi:10.1109/ICASSP.2017.7953152.

[22] XU, H., K. LI, Y. WANG, J. WANG, S. KANG, X. CHEN, D. POVEY, and S. KHUDANPUR: *Neural network language modeling with letter-based features and importance sampling*. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6109–6113. IEEE, 2018. doi:10.1109/ICASSP.2018.8461704.

[23] BERNARD, M. and H. TITEUX: *Phonemizer: Text to phones transcription for multiple languages in python*. *Journal of Open Source Software*, 6(68), p. 3958, 2021. doi:10.21105/joss.03958. URL `https://doi.org/10.21105/joss.03958`.

[24] VAN BEZOOIJEN, R. and R. VAN DEN BERG: *Word intelligibility of language varieties in the Netherlands and Flanders under minimal conditions*. *Linguistics in the Netherlands*, 16(1), pp. 1–12, 1999. doi:10.1075/avt.16.03bez.

[25] IMPE, L. and D. GEERAERTS: *Babel in Vlaanderen - Een experimenteel onderzoek naar onderlinge verstaanbaarheid tussen Vlamingen*. *Studies van de BKL*, 3, 2008.

[26] IMPE, L., D. GEERAERTS, and D. SPEELMAN: *Mutual intelligibility of standard and regional Dutch language varieties*. *International Journal of Humanities and Arts Computing*, 2(1-2), pp. 101–117, 2008. doi:https://doi.org/10.3366/E1753854809000330.

[27] VAN DE VELDE, H.: *Variatie en verandering in het gesproken Standaard-Nederlands (1935-1993)*. Phd thesis, Katholieke Universiteit Nijmegen, Nijmegen, 1996. URL `https://repository.ubn.ru.nl/handle/2066/146159`.

[28] GARNERIN, M., S. ROSSATO, and L. BESACIER: *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*, AI4TV '19, pp. 3–9. Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3347449.3357480. URL `https://doi.org/10.1145/3347449.3357480`.

[29] ADDA-DECKER, M. and L. LAMEL: *Do speech recognizers prefer female speakers?* In *Proceedings of Interspeech 2005*, pp. 2205–2208. 2005. doi:10.21437/Interspeech.2005-699.