A Comparative Study on the Biases of Age, Gender, Dialects, and L2 speakers of Automatic Speech Recognition for Korean Language

Jonghwan Na*, Yeseul Park† and Bowon Lee‡
* Inha University, Incheon
E-mail: jhna@dsp.inha.ac.kr Tel/Fax: +82 10-4623-0591
† Inha University, Incheon
E-mail: yspark@dsp.inha.ac.kr Tel/Fax: +82 10-2587-8450
‡ Inha University, Incheon
E-mail: bowon.lee@inha.ac.kr Tel/Fax: +82 32-860-7423

***Abstract*—Recent advancements in the field of Automatic Speech Recognition (ASR) have seen the emergence of large-scale models, contributing to a surge in research and development. The performance of recent ASR models has rapidly improved with the utilization of extensive pre-training datasets. However, challenges persist in enhancing the recognition accuracy for non-mainstream groups such as the elderly and speakers of regional dialects. This paper conducts experiments using Korean speech data to compare and analyze the biases related to gender, age, dialects, and second language (L2) Korean speakers using the Conformer, wav2vec 2.0, and Whisper models. The experimental results showed that female results exhibited better performance in ASR models than those of males, and Whisper exhibited lower biases than two other models in most cases. Furthermore, Whisper demonstrated robustness compared to the other two models in the L2 speakers. Additionally, the analysis of characters with high error rates for each group revealed that, in the case of Korean, spacing and particles exhibited high error rates. It was also observed that characters with high error rates were similar within age groups rather than between gender groups. In this study, we conducted the first-ever examination of various biases in Korean ASR. The identified biases through these experiments may serve as a starting point for research aimed at improving the performance of ASR for non-mainstream groups. This study underscores the significance of addressing biases to advance fairness in the field of ASR.**

I. Introduction

Since the introduction of the transformer architecture [1], various models based on transformers have been proposed in the field of Automatic Speech Recognition (ASR), and their performances have significantly improved. Under these circumstances, research efforts in the field of ASR are actively focused on enhancing inclusiveness and accessibility [2]–[8]. As part of such research, studies on biases in ASR performance across different cohorts is being conducted from various perspectives, such as studies analyzing bias across different age groups [7], [9], between genders [7], [9]–[14], and based on racial backgrounds [4], [7], [12], [15]. Studies on such biases are not limited to English [7], [12]–[15] but are also extensively conducted based on various languages, including Dutch [9], [16], [17], French [10], [13], Mandarin [11], and Arabic [18].

According to prior research, gender bias in ASR models generally indicates that these models exhibit higher performance for females than males [9], [11], [13]. However, some studies have reported instances where they demonstrate superior performance for males than females [10], [14]. Regarding age-based bias, Feng et al. reported that, on a Dutch dataset, higher performance is observed in the order of teenagers, older individuals, and young children [9]. Conversely, Liu et al. reported superior performance in the elderly compared to other age groups on the English dataset [7]. Thus, the bias in ASR models appears to vary depending on the language used and the specific model employed in the experiments. To the best of our knowledge, no study comparing and analyzing biases of ASR models on Korean language has been reported in the literature.

In this regard, we conducted experiments in three different cases of ASR training approaches: training a model from scratch, fine-tuning of a model pre-trained through self-supervised learning, and finally, fine-tuning of a model pre-trained through supervised learning using three different ASR models, namely Conformer [19], wav2vec 2.0 [20], Whisper [21], respectively. We evaluated the performance of these ASR models for three age groups – children, adults, and elderly – and two genders, male and female. We examined and compared biases within each group. Additionally, we investigated the regional biases of six dialects of Korea, namely, the dialects of the greater Seoul, the Korea's metropolitan area (MA), Chungcheong (CC), Gyeongsang (GS), Jeolla (JL), Gangwon (GW), and Jeju (JJ). Finally, we compared the speakers whose native languages are English, Chinese, and Japanese, learning Korean as a second language (L2).

The identified biases through these experiments may potentially serve as a starting point for research aimed at reducing the bias of ASR especially for non-mainstream groups including children, elderly individuals, speakers of regional dialects, and L2 speakers. This study underscores the importance of analyzing biases to advance fairness in the field of ASR.

TABLE I: Number of speakers, number of utterances, duration, and the average number of characters in train and test sets

Group	Number of speakers			Number of utterances			Duration (hours)			Avg. number of characters		
	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
train-children	105	135	240	50,000	50,000	100,000	65	62	127	14.23	14.17	14.20
train-adults	465	775	1,240	50,000	50,000	100,000	68	73	141	19.33	19.36	19.35
train-elderly	462	495	957	50,000	50,000	100,000	123	126	249	35.67	35.19	35.43
train-total	1,032	1,405	2,437	150,000	150,000	300,000	256	261	517	23.08	22.91	22.99
test-children	56	72	128	30,000	30,000	60,000	38	38	76	14.31	13.98	14.15
test-adults	273	361	634	30,000	30,000	60,000	39	40	79	19.26	19.52	19.39
test-elderly	245	260	605	30,000	30,000	60,000	74	77	151	36.47	36.01	36.24
test-MA	25	29	54	12,083	13,806	25,889	25	28	53	26.68	26.84	26.77
test-CC	7	8	15	4,588	5,026	9,614	8	9	17	25.92	25.86	25.89
test-GS	9	12	21	4,203	6,449	10,652	8	10	18	28.73	29.02	28.91
test-JL	4	5	9	2,160	2,971	5,131	4	5	9	26.56	26.63	26.60
test-GW	3	3	6	1,637	1,481	3,118	3	2	5	26.35	26.77	26.55
test-JJ	1	1	2	555	512	1,067	1	1	2	28.81	29.14	28.97
test-english L2	37	107	144	673	1,994	2,667	3	7	10	57.16	57.14	57.15
test-chinese L2	32	311	343	1,732	1,828	3,560	6	6	12	57.04	57.20	57.12
test-japanese L2	13	215	228	721	1,847	2,568	2	6	8	57.22	56.29	56.55

II. Experimental Setup

A. Datasets

Table I presents the statistics of the training and testing datasets used in the experiments. The datasets employed in the experiments were sourced from AI-hub[1]. Four distinct types of datasets, namely "Free Conversation Speech (Infants and Children)", "Free Conversation Speech (General Adults)", "Free Conversation Speech (Elderly)", and "Speech Data of Non-Native Korean Speakers for AI Training", were chosen and employed in the experiments. The training data utilized in the experiments were randomly sampled from the remaining three datasets, excluding non-native speakers, with $100,000$ utterances selected from each dataset. In this way, a total of $300,000$ utterances comprised the training data.

Throughout this process, the gender ratio within the training dataset was adjusted to achieve the of ratio of 1:1, aligning with the objectives of the experiments. The training dataset comprises a total duration of 517 hours, and the number of utterances was uniformly distributed across all groups. Examining the average number of characters in transcripts reveals that, in descending order, the elderly, adults, and children exhibit higher counts. The test datasets were uniformly composed of $30,000$ utterances for each age and gender group, and the regional and L2 speaker datasets were obtained as described in Table I. These utterances were randomly extracted for each age and gender category, ensuring the exclusion of shared speakers between the training and test datasets.

B. Character Error Rate

For experiments, the Character Error Rate (CER) is adopted as the evaluation metric to conform with the common practices for Korean ASR due to the unique characteristics of the language. The equation for calculating the CER is

$$CER = \frac{S + D + I}{N}, \tag{1}$$

where it is computed by summing the number of deletions (D), substitutions (S), and insertions (I) of the characters, and then dividing it by the total number of characters (N).

C. Models

We selected three ASR models – Conformer [19], wav2vec 2.0 [20], and Whisper [21] – in our experiments. The reason for selecting the Conformer was to analyze biases in a model trained from scratch. The model based on Conformer is comprised of a Conformer encoder and a Connectionist Temporal Classification (CTC) [22] decoder. The Conformer encoder is configured as Conformer-large with 17 layers and 8 attention heads, amounting to a total of 118.8 million parameters. The initial learning rate was set to 1e-0.6, with a batch size of 32. The experiment utilized a tri-stage learning rate scheduler, and the optimizer employed was Adam.

The rationale behind the selection of wav2vec 2.0 was driven by the need to examine the bias inherent in pre-trained models when utilizing unlabeled data. Like the Conformer, wav2vec 2.0 was a model that demonstrated the SOTA performance upon its release. For the experiment, the encoder was configured as wav2vec 2.0 while the decoder was set as CTC. The wav2vec 2.0 encoder utilized a large model consisting of 24 transformer blocks and attention heads. Additionally, the experiment utilized the pre-trained wav2vec 2.0 XLSR-53 model, trained on a $56,000$-hour dataset spanning 53 languages. The wav2vec 2.0 XLSR model contains 300 million parameters, with an initial learning rate of 4e-4, a batch size of 32, and a linear learning rate scheduler. Adam optimizer was used for the experiments.

Whisper is a pre-trained model like wav2vec 2.0 except it is pre-trained with labeled data. This distinction led to its

[1] https://aihub.or.kr/

inclusion in the experiments to scrutinize biases in models pre-trained through supervised learning on multilingual data. Both its encoder and decoder are structures with transformer-based layers. For the experiments detailed in this paper, the Whisper-small model was utilized, comprising 12 layers and 12 attention heads, totaling 244 million parameters. The initial learning rate was configured at 5e-07, accompanied by a batch size of 32, and a linear learning rate scheduler was employed. The Adam optimizer was used for training. Finally, all models were trained for 25 epochs, incorporating an early stopping mechanism.

III. Results and Discussions

A. Age and Gender Biases

Table II shows the CERs on the test datasets of the different age groups and gender groups. In terms of overall performance, it can be observed that wav2vec 2.0, Whisper, and Conformer exhibit superior performance in the order mentioned, yielding the average CERs of 2.15%, 3.24%, and 5.17% respectively.

When comparing performance based on gender, it can be observed that across all models, the test data for females exhibit slightly better performance than males. This suggests an ASR models have a bias favoring females in Korean, regardless of age. Conformer demonstrated the most pronounced gender-based bias, with a difference in CER values between genders being 1.02. In contrast, wav2vec 2.0 and Whisper demonstrated higher resilience to gender-based bias, with values of 0.43 and 0.71, respectively. These results imply that models pre-trained on various and extensive datasets, such as wav2vec 2.0 and Whisper, are less susceptible to biases compared to the model trained from scratch. It is interesting to observe that including multi-lingual dataset helps the performance of the target language.

Upon comparing performance across different age groups, Conformer demonstrated better performance in the order of adults, children, and elderly testsets. For wav2vec 2.0, the best performance was observed in the order of children, adults and the elderly. Lastly, for Whisper, performance was found to be superior in the order of adults and elderly, with an interesting observation that performance on the children's dataset was the lowest. The reason why Whisper exhibited the lowest performance on the children test dataset, unlike the other models, is likely due to the fact that the Whisper model takes the hyperparameter '*max length*' as input. This is speculated to result in a higher CER for the relatively shorter transcript length, with hallucination issues occurring in some cases for Whisper [23]. In this experiment, as can be seen in Table I, the sentences of children dataset is shorter than those of the other two datasets. The age-based standard deviations of CER are 0.70 for Whisper, 0.78 for Conformer, and 0.98 for wav2vec 2.0, indicating that Whisper shows the smallest bias.

Table III displays the proportions of the sentences for non-zero CERs and Fig. 1 displays the distribution of their CER values by age and gender across three models. As evident from Table III, it is apparent that the model with the highest performance, wav2vec 2.0, exhibits the lowest proportion of non-zero CER. Comparing non-zero CER across genders reveals that females consistently exhibit lower values than males across all cases, consistent with the overall trends observed in Table II. Moreover, the gender disparity in non-zero CER proportions is also least pronounced in wav2vec 2.0. An intriguing observation is that while Table II shows a marginal difference in performance between wav2vec 2.0 and Whisper on the elderly test dataset, Table III indicates that Whisper has a lower proportion of non-zero CER, suggesting that although Whisper may correctly infer more sentences, the average CER of erroneous sentences is higher compared to wav2vec 2.0.

TABLE II: CER (%) on Age and Gender-Specific Test Datasets using Conformer large, wav2vec 2.0, and Whisper-small

Model	Children		Adults		Elderly	
	Male	Female	Male	Female	Male	Female
Conformer	4.98	5.14	4.96	3.93	7.09	4.92
wav2vec 2.0	1.57	1.56	1.88	1.57	3.64	2.65
Whisper-small	4.32	3.57	2.85	2.25	3.63	2.84

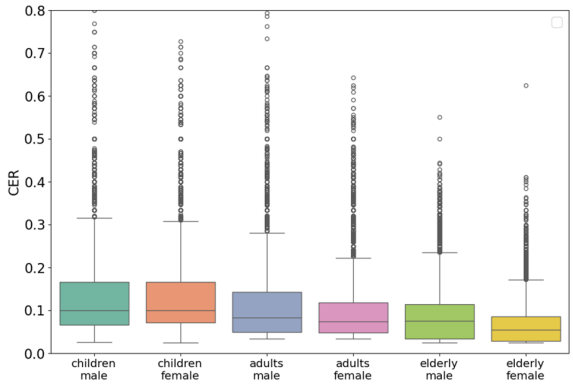
TABLE III: Proportion of non-zero CER sentences in the test set

Model	Children		Adults		Elderly	
	Male	Female	Male	Female	Male	Female
Conformer	0.45	0.45	0.49	0.44	0.82	0.73
wav2vec 2.0	0.15	0.15	0.22	0.19	0.67	0.58
Whisper-small	0.38	0.29	0.29	0.26	0.60	0.52

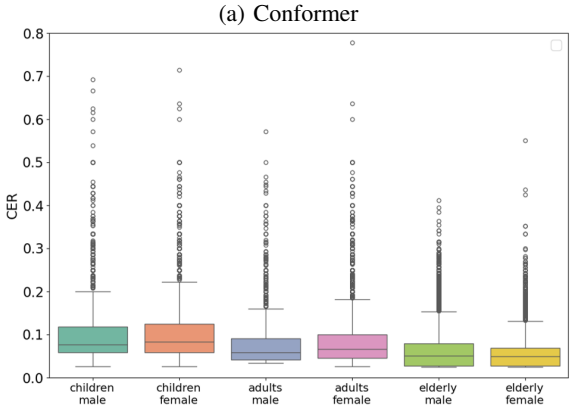
In Fig. 1, for the model with the best overall performance, wav2vec 2.0, the outliers in non-zero CER values are relatively lower compared to the other two models. Additionally, examining the distribution of outliers by gender in Fig. 1 reveals that, paradoxically, in cases where performance was better for females, there are more instances of higher outlier values. Viewing Fig. 1 alongside Table III, it becomes apparent that Whisper exhibits instances where the CER exceeds 0.5 for male and female children, who have the shortest transcripts. Conversely, Table III indicates that Whisper has a lower non-zero CER than the other two models for the elderly, who have the longest transcripts. This is considered as a result arising from the inherent characteristics of the model, as mentioned earlier.

TABLE IV: CER (%) on Korean regional-specific test datasets using Conformer large, wav2vec 2.0, and Whisper-small

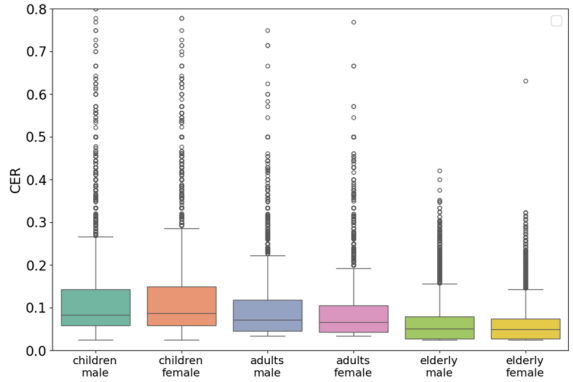
Model	**MA**	**CC**	**GS**	**JL**	**GW**	**JJ**	**Avg**	**SD**
Conformer	5.02	6.25	6.97	4.08	7.96	9.14	6.57	1.87
wav2vec 2.0	2.67	3.10	3.82	2.18	4.14	5.03	3.49	1.04
Whisper-small	2.92	3.52	4.17	2.37	4.23	4.85	3.68	0.92



(a) Conformer



(b) wav2vec 2.0



(c) Whisper-small

Fig. 1: Distributions of non-zero Character Error Rates for Conformer, wav2vec 2.0, and Whisper-small

B. Regional and L2 Biases

The experimental results with six different regions in Korea are summarized in Table IV. It was evident that the performance of the test set from the JL region was the best across all models. Conversely, the performance on the JJ dataset was consistently the lowest across all models. The distinctiveness of the JJ dialect due to its geographical isolation as an island likely contribute to the highest CERs. This is also supported by studies showing that JJ dialect is largely incomprehensible to monolingual speakers of standard Korean, indicating a significant linguistic divide [24]. Examining the

TABLE V: CER (%) on L2 speakers of English, Chinese, and Japanese Test Datasets using Conformer large, wav2vec 2.0, and Whisper-small

Model	English			Chinese			Japanese		
	Male	Female	All	Male	Female	All	Male	Female	All
Conformer	21.72	19.49	20.09	22.74	19.99	21.33	21.65	19.36	20.02
wav2vec 2.0	17.95	16.54	16.89	19.05	17.08	18.03	18.67	17.09	17.52
Whisper-small	10.10	9.96	10.00	13.17	12.00	12.57	12.80	12.59	12.65

average performance for the six regions, wav2vec 2.0 achieved the best performance with a CER of 3.49%, followed by Whisper and Conformer with CERs of 3.68% and 6.57%, respectively. Checking the standard deviation of CERs for the six regions, Whisper exhibited the lowest value at 0.92, indicating that Whisper has the least bias towards regional dialects. This suggests that Whisper, being pre-trained on diverse languages, and datasets, is more robust to biases.

Table V shows the experimental results with three different Korean L2 speakers whose native languages are English, Chinese, and Japanese. The results are further broken down by gender, with separate CER values for male, female, and all speakers combined. The experimental results indicated that English native L2 speakers achieved the highest performance in Korean speech recognition, followed by Japanese and Chinese native L2 speakers. Analyzing gender bias among L2 groups, Chinese native L2 speakers exhibit the most significant gender bias in CER with an average bias of 1.96, while English native L2 speakers show the least gender bias with an average bias of 1.26. Among the three models, Whisper consistently demonstrated superior performance across all language groups and genders, achieving the lowest CERs compared to Conformer and wav2vec 2.0. This suggests that Whisper may be more effective in handling the speech variability of L2 speakers, providing more accurate speech recognition capabilities. The results also highlight a consistent pattern where female speakers tend to have slightly lower CERs compared to male speakers across all models and languages.

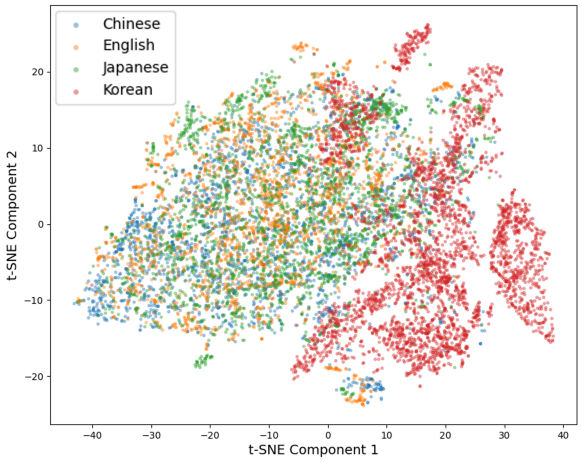


Fig. 2: Distributions of t-SNE of MFCC features for L2 speakers from China, Japan, English-speaking countries, and native Korean speakers

TABLE VI: Top 5 error characters excluding 9 consistently high-CER characters (/ /, /ʌ/, /nɯn/, /ɛ/, /hɛ/, /i/, /ɯn/, /kʌ/, /ɯl/) across models, by gender and age group

		Conformer	wav2vec 2.0	Whisper-small
Children	M	/nɑ/, /ɑ/, /kɑ/, /mjʌn/, /hɑ/	/kɑ/, /ɑ/, /hɑ/, /nɑ/, /il/	/nɑ/, /ɑ/, /kɑ/, /hɑ/, /ko/
	F	/kɑ/, /nɑ/, /ɑ/, /to/, /il/	/kɑ/, /ɑ/, /to/, /it/, /ko/	/kɑ/, /ɑ/, /nɑ/, /ko/, /to/
Adults	M	/hɑ/, /ɑ/, /nɑ/, /ja/, /tɕi/	/ko/, /kɑ/, /ku/, /to/, /hɑ/	/ku/, /ɯi/, /hɑ/, /ko/, /ɑ/
	F	/ɯi/, /nɑ/, /ɑ/, /hɑ/, /tɕi/	/ko/, /ɯi/, /ɑ/, /ku/, /hɑ/	/ku/, /ɯi/, /hɑ/, /ɑ/, /nɛ/
Elderly	M	/hɑ/, /kɑ/, /sʌ/, /nɑ/, /ɯi/	/hɑ/, /kɑ/, /ɯi/, /lɯl/, /ko/	/kɑ/, /hɑ/, /sʌ/, /ɯi/, /ko/
	F	/kʌt/, /kɑ/, /hɑ/, /ɯi/, /sʌ/	/ɯi/, /kʌt/, /kɑ/, /hɑ/, /ɑ/	/kʌt/, /hɑ/, /ɑ/, /kɑ/, /nɛ/

Figure 2 shows the t-SNE plot of Mel-frequency cepstral coefficients (MFCC) features extracted from Korean L2 speakers whose native languages are Chinese, English, and Japanese, as well as native Korean speakers. Notably, there is a significant distinction between native Korean speakers and the other L2 speaker groups. The Chinese L2 speaker cluster exhibits a wider distribution, likely due to the presence of tonal characteristics in the Chinese language. Additionally, the t-SNE distributions of Japanese and English L2 speakers show relatively overlapping regions compared to the other languages. These differences indicate clustering patterns that reflect the underlying acoustic and phonetic influences of the speakers' native languages.

C. Characters with High CER

To analyze the bias results of each group, we examined characters exhibiting high CERs. The investigation revealed that among the top 20 characters with the highest errors for each gender and age group across three models, 9 characters – / /, /ʌ/, /nɯn/, /ɛ/, /hɛ/, /i/, /ɯn/, /kʌ/, /ɯl/ – appeared consistently. Notably, the character '/ /' representing the space between words, was the most error-prone in all cases. Also, the characters corresponding to Korean particles, which are functional elements attached to nouns, verbs, or adjectives to indicate grammatical relations, were frequently misinferred. This is considered to be influenced by the characteristics of the Korean language, attributed to the somewhat intricate rules of spacing and the flexible utilization of particles, which are well known to be two major sources of grammatical errors made even by humans [25], [26]. Table VI presents the top 5 characters, excluding the aforementioned 9 characters, that exert the highest influence on CER for each model across gender and age groups. The top error characters exhibit a high degree of similarity between males and females within the same age group although they differ among the age groups.

IV. Conclusion and Future Works

This paper presented a comparative analysis of various biases for Korean ASR, and the experimental findings provide insights into the performance and biases of the recent ASR models trained with Korean speech datasets. Wav2vec 2.0 and Whisper boast notable robustness to biases, especially regarding the gender difference. The analysis revealed a consistent bias favoring females across all models, with Conformer exhibiting the most pronounced gender-based bias. Furthermore, we observed that age-based bias is least prominent in the adults category across all models, with Whisper exhibiting the smallest age-based bias. Regional biases were also examined, revealing that the Whisper model demonstrates the least amount of regional bias, potentially by training with the vast amount of multi-lingual datasets, and all models exhibit the lowest CER in the JL region. Through experiments, the performance differences of Korean ASR models for L2 speakers were also examined, revealing that the models performed best with English native speakers, followed by Japanese and Chinese.

Additionally, it was observed that the Whisper model outperformed the other two models, which can be attributed to the fact that Whisper was pre-trained on a large amount of labeled data. This pre-training likely contributed to its robustness against various speech patterns and contextual biases in transcriptions.

In line with previous studies of other languages, we have substantiated the presence of biases in Korean ASR models through our experiments. If these biases in ASR models can be addressed properly, it can lead to the development of more inclusive ASR models, which could result in overall performance improvements.

Research on analyzing the identified biases from acoustic, prosodic, and phonetic perspectives through experiments and exploring methods to resolve them is planned as further research. Acquiring additional dialectal datasets will be desirable for more thorough examination of the regional biases. Also, considering the influence of native languages is crucial when evaluating and improving ASR model performance for L2 speakers. Therefore, understanding the clear differences between native languages is deemed necessary. Hence, it is believed that further investigation into the differences among L2 speakers from various perspectives will be required in future studies.

Acknowledgment

This work was supported by the Ministry of Science and ICT of the Republic of Korea and the National Research Foundation of Korea (NRF-2023R1A2C2006725) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) (RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] T. Wang, S. Hu, J. Deng, *et al.*, "Hyper-parameter Adaptation of Conformer ASR Systems for Elderly and Dysarthric Speech Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1733–1737.

[3] P. DHERAM, M. Ramakrishnan, A. Raju, *et al.*, "Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities," in *Proc. Interspeech 2022*, 2022, pp. 1268–1272.

[4] J. L. Martin and K. Tang, "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"," in *Proc. Interspeech 2020*, 2020, pp. 626–630.

[5] R. Gale, L. Chen, J. Dolata, J. van Santen, and M. Asgari, "Improving ASR Systems for Children with Autism and Language Impairment Using Domain-Focused DNN Transfer Techniques," in *Proc. Interspeech 2019*, 2019, pp. 11–15.

[6] M. Moore, H. Venkateswara, and S. Panchanathan, "Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems," in *Proc. Interspeech 2018*, 2018, pp. 466–470.

[7] C. Liu, M. Picheny, L. Sarı, *et al.*, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6162–6166.

[8] H. Gao, X. Wang, S. Kang, *et al.*, "Seamless equal accuracy ratio for inclusive ctc speech recognition," *Speech Communication*, vol. 136, pp. 76–83, 2022.

[9] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[10] M. Garnerin, 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*, 2019, pp. 3–9.

[11] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101 567, 2024.

[12] 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.

[13] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" In *Ninth European Conference on Speech Communication and Technology*, 2005.

[14] R. Tatman, "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 53–59.

[15] M. P. Y. Chan, J. Choe, A. Li, Y. Chen, X. Gao, and N. Holliday, "Training and typological bias in ASR performance for world Englishes," in *Proc. Interspeech 2022*, 2022, pp. 1273–1277.

[16] 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)*, IEEE, 2023, pp. 146–151.

[17] Y. Zhang, A. Herygers, T. Patel, Z. Yue, and O. Scharenborg, "Exploring data augmentation in bias mitigation against non-native-accented speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–8.

[18] G. Droua-Hamdani, S.-A. Selouani, and M. Boudraa, "Speaker-independent asr for modern standard arabic: Effect of regional accents," *International Journal of Speech Technology*, vol. 15, pp. 487–493, 2012.

[19] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[23] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane, "Careless whisper: Speech-to-text hallucination harms," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1672–1681.

[24] C. Yang, W. O'Grady, S. Yang, N. H. Hilton, S.-G. Kang, and S.-Y. Kim, "Revising the language map of korea," *Handbook of the Changing World Language Map*, pp. 215–229, 2020.

[25] D.-J. Kim, "A method for detection and correction of pseudo-semantic errors due to typographical errors," *Journal of the Korea Society of Computer and Information*, vol. 18, no. 10, pp. 173–182, 2013.

[26] S.-H. Lee, M. Dickinson, and R. Israel, "Developing learner corpus annotation for korean particle errors," in *Proceedings of the Sixth Linguistic Annotation Workshop*, 2012, pp. 129–133.