Exploring the Relationship Between Bias and Speech Acoustics in Automatic Speech Recognition Systems

An Experimental Investigation Using Acoustic Embeddings and Bias Metrics on a Dataset of Spoken Dutch

Piotr Cichoń[1]

Supervisors: Odette Scharenborg[1]**, Jorge Martinez Castaneda**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Piotr Cichoń
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, Jorge Martinez Castaneda, Merve Gürel

Abstract

Automatic Speech Recognition (ASR) systems have become an integral part of daily lives. Despite their widespread use, these systems can exhibit biases that express themselves in the differences in their accuracy and performance across different demographic groups. Methods quantifying these biases have been developed. This paper investigates the relationship between bias and the acoustic characteristics of speakers. By examining various acoustic embeddings, derived from models like wav2vec 2.0 and XLSR, we aim to identify which embeddings correlate most strongly with bias. The findings offer insights into improving the fairness of ASRs by exploring how acoustic features influence bias in ASR systems. Future research directions include exploring isolated speech properties and extending the study to diverse linguistic contexts to deepen understanding in this area.

1 Introduction

Automatic Speech Recognition (ASR) systems are systems that convert speech to text. They operate through several stages: capturing audio, extracting features, and using acoustic and language models to decode phonetic sequences into text. These systems are becoming increasingly common in our daily lives.

It is, therefore, critical to ensure that these systems are working properly. The performance of ASR systems is typically measured by their speech recognition accuracy. Unfortunately, it is not uncommon for the ASR systems to be biased against certain groups of people [11, 12]. Despite the system being overall highly accurate, disparities in recognizing speech from different groups can lead to a worse user experience for specific demographics. For example, the ASR systems tend to recognise the speech with a different accuracy depending on the race [14], gender [2], or the age [2] of the speaker.

Bias can have numerous different origins. The quality of the data points in the training dataset, how varied its composition is, or even the diversity of the developer team can all be the reasons for the existence of bias [12].

The direct cause where the bias occurs often lies in how the ASR system processes its input. Factors such as speaker vocabulary and acoustics significantly impact ASR performance. Speech acoustics mirror the structure and characteristics of the vocals that generated them [19]. Here, we refer to acoustics as the features extracted from the speech signal that allow for distinguishing between speakers. The main research question that this paper answers is:

- *How are the bias of an ASR system and the acoustics of the speaker related?*

It is unclear what features distinguish different types of speech best. Different features are compared in order to find a suitable one for a given task. For example, research concerning the selection of a set of features that would best reflect the emotions of a speaker has been ongoing for many years [3, 13]. In this paper, we refer to these features as acoustic embedding. Instead of emotions, we aim to find an acoustic embedding that reflects the bias experienced by the speaker. Therefore, our secondary research question is:

- *Which acoustic embedding best reflects the bias?*

2 Background and related work

The two main components under investigation in this paper are bias and acoustic embeddings. Research has been done on their relationship with other factors but not with each other [5, 6, 11, 12].

Previous studies have analyzed the relationship between bias and phonemes, revealing that certain phonemes are more prone to misrecognition, which can contribute to bias against specific groups of speakers [11, 12]. By examining phoneme error rates (PER), researchers have identified atypical pronunciations as a potential source of bias.

Next to the bias, acoustic embeddings play an important role in this research. They are numerical representations derived from speech signals encapsulating essential acoustic features. They represent speech characteristics such as intonation, pitch, energy distribution, and phonetic content. The idea is similar to textual word embeddings, which create similar vector representations for words with similar meanings. However, acoustic embeddings are designed to capture acoustic similarities rather than semantic ones. In the acoustic embedding space, the goal is to organize speech in a way that groups together similar sounds.

Previous studies investigated the relationship between acoustic embeddings and native-likeness ratings. These ratings are assessments by native speakers on how closely speech resembles native speech patterns. Acoustic embeddings derived from methods like Fourier transforms and neural networks, such as wav2vec 2.0 and XLSR, have shown significant correlations with these ratings [5, 6].

In this paper, we explore how the bias of the ASR system is related to the variability in the acoustics of the speech captured by acoustic embeddings. By examining the relationship between the differences in these embeddings and bias, it is possible to find how variations in the acoustic embeddings could reflect underlying biases in the ASR system's performance across different speaker groups.

3 Methodology

This section outlines the methodology of an experimental study that is conducted to answer the research questions. This section describes the dataset, the design choices, and the experiment. Figure 1 depicts the workflow of the experiment. The starting point from which all the components are derived are the speech files, transcriptions, and the metadata of the speakers, which constitute the dataset used. The following steps involve deriving the acoustic-based embeddings, calculating the distances between them, and calculating the bias. Finally, a method to relate the acoustic-based distances and bias is described, which is the focus of this paper.

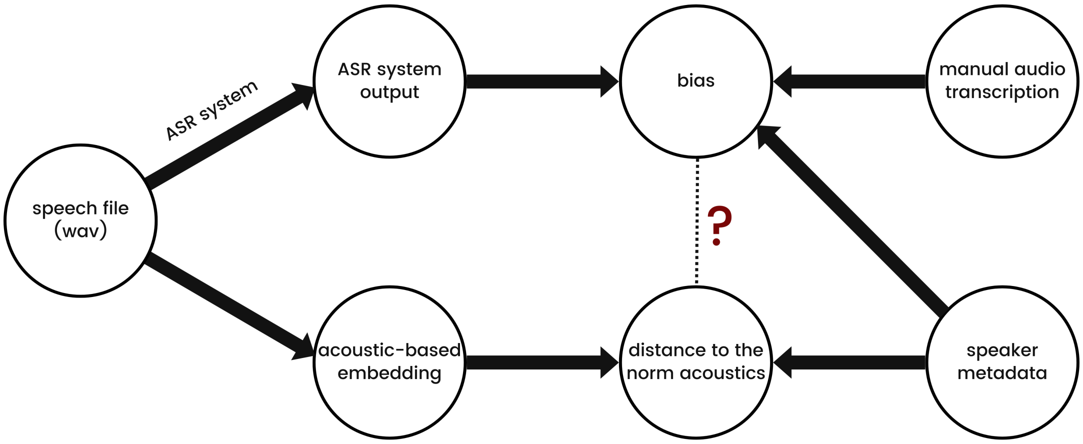

Figure 1: Diagram illustrating the experiment procedure. The arrows indicate the derivation order, and the question mark symbolises relating speaker acoustics to bias.

Datasets

The Spoken Dutch Corpus (CGN) is the dataset that was by Patel et al [17] for training the ASR systems for which the bias is calculated (ASR system in Figure 1). The dataset contains 900 hours of spoken Dutch which amounts to about 9 million words. The speakers come from different regions of the Netherlands as well as the Flanders region in Belgium and are aged 18-65 years. There are a total of 1185 female and 1678 male speakers who speak in various styles, from non-staged casual talk to reading. In this study, we build on the information derived from the outputs of ASR systems found by Patel et al [17].

The Jasmin-CGN [9] is the dataset used for the experiments. It consists of the speech files, manual audio transcriptions, and speaker metadata. It contains two types of speech: read (Read) and human-machine interaction speech (HMI). In this study, we use a subset of the dataset focusing on Dutch-native children aged 7 to 16 years. The group includes 52 males and 48 females.

ASR system

The study employs two primary ASR models: a Conformer model and the OpenAI Whisper small model. The Conformer model, a state-of-the-art neural network for speech recognition, is available in three configurations: NoAug, SpAug, and SpSpecAug. The second model is the OpenAI Whisper, which comes in two configurations: Ws and WsFTcgn. For detailed configuration specifics of these models, readers are encouraged to refer to the paper by Patel et al [17]. In the experiments, the word error rates (WERs) based on the output of these systems, as calculated by the author cited in this paper, are utilized.

Ground truth for the bias

The bias metrics aim to capture the performance differences of an ASR between different speakers. The performance is measured by how similar the output is to the reference human-made annotation of what was said. Ideally, the ASR system would produce identical text to the annotation. The similarity between the annotation and the output of the system is measured using the word error rate (WER). A low WER indicates high similarity, while a high WER indicates greater dissimilarity. The calculation of WER is as follows:

$$\text{WER} = \frac{S + I + D}{W} \tag{1}$$

where:

$$\begin{aligned} S &= \text{the number of substitutions} \\ I &= \text{the number of insertions} \\ D &= \text{the number of deletions} \\ W &= \text{the number of words in the reference text} \end{aligned}$$

Substitution takes place when a word in the result is replaced by another word that does not exist in the reference text. For example, if the reference text states ”It is sunny” and the ASR output transcribes it as ”It is cloudy”, this constitutes a substitution error. Insertions include additional words that appear in the ASR results but not in the reference text, e.g. recognising ”red apple” instead of ”apple” in effect adding the word ”red”. On the other hand, deletion occurs when a word from the reference text is missing in the ASR results, e.g., ”mobile phone” is interpreted as ”phone” and the word ”mobile” is omitted.

The bias metrics used are those proposed by Patel et al [17]:

- group-to-min absolute:

$$\text{bias}(spk) = \text{WER}_{spk} - \text{WER}_{\text{min}} \tag{2}$$

- group-to-norm absolute:

$$\text{bias}(spk) = \text{WER}_{spk} - \text{WER}_{\text{norm}} \tag{3}$$

- group-to-min relative:

$$\text{bias}(spk) = \frac{\text{WER}_{spk} - \text{WER}_{\text{min}}}{\text{WER}_{\text{min}}} \tag{4}$$

- group-to-norm relative:

$$\text{bias}(spk) = \frac{\text{WER}_{spk} - \text{WER}_{\text{norm}}}{\text{WER}_{\text{norm}}} \tag{5}$$

There are two different baselines these metrics utilise: comparing the WER to the WER of the normative (norm) group (equations 3 and 5) or the speaker group with the lowest WER (equations 2 and 4).

For this research, we opted for the metric that uses min group because it can be calculated using a dataset that lacks a norm group, thereby enhancing reproducibility. We chose the absolute (abs) metric because of its simplicity and intuitive understanding. The chosen metric is, therefore, the one presented in equation 2.

Acoustic-based embeddings

Given the uncertainty regarding the relationship between acoustic variability and bias size, different ways of capturing the acoustic variability are considered. These representations are in the form of sequences of numbers that capture acoustic properties from speech. We present four candidates, of which two are chosen for the experiments.

I-vectors are fixed-length low-level representations of speech that are meant to be used in tasks such as speaker verification [10]. They are derived from features like Mel-Frequency Cepstral Coefficients (MFCCs), which capture the spectral properties of speech. These properties are, for example, pitch peaks or smoothness of the sound. However, a study on language characterisation has shown that the i-vectors have the highest correlation with syntax differences but do not correlate with phonology or the phoneme inventory [18]. This limitation suggests that i-vectors might not be suitable for capturing the differences in the speech across different ages, for example. The authors of a study on speech differences between 5 and 10-year-old children and adults suggest that although children have developed the phoneme inventory, the phonological aspects might not be fully developed yet at that age [15]. For this reason, the i-vector representation is not used in the experiments.

An embedding method proposed by Bartelds et al. [6], for a large part, builds on Mel-Frequency Cepstral Coefficients (MFCCs). Apart from MFCCs, it augments the energy derived from the speech. In the same study, it was found that distances between these representations strongly correlate (absolute value of 0.71) with how natural or native-like the speech sounds. In other words, the closer the distances between these representations are, the more closely the speech is perceived by listeners as resembling natural or native speech. This correlation suggests that their method effectively captures features in the speech signal that are important for producing speech that sounds natural or native-like to the ears of a native speaker. Although this correlation value is considered strong, two embedding methods achieved a higher one, and they will used in our experiments.

The model wav2vec 2.0 (w2v2) [4] is a self-supervised speech representation model developed by Facebook AI. It converts raw audio into latent speech representations in a way inspired by methods used in natural language processing. The model was pre-trained on the large, unlabeled Librispeech dataset [16], which includes 960 hours of English speech. The model's architecture consists of multiple hidden layers that capture information about the speech signal. Although the specific nature of these encoded properties is not explicitly defined, research has demonstrated a strong correlation (absolute value of 0.86) between w2v2's representations and native-likeness, highlighting its capability to capture essential features of speech [5].

Cross-Lingual Speech Representation (XLSR) [8] is another self-supervised model that uses neural networks. However, in contrast to w2v2, it is designed to generalize to multiple languages. It was trained on 56,000 hours of speech from 53 languages to achieve this. This model correlated with the absolute value of 0.78 to the native-likeness labels [5]. Along with w2v2, the model achieves the highest correlation with native-likeness labels and is therefore incorporated in the experiments.

However, these two embedding methods require training in contrast to the one based on MFCC. The trained models fine-tuned for Dutch for both w2v2 and XLSR are publicly available [7] and are used in this research. It is worth noting that they have been trained on a dataset different from Jasmin-CGN so they do not overfit on the data that is used in the experiments.

The output produced from the speech of a speaker as input by w2v2 or XLSR is the *acoustic embedding*. The acoustic embeddings are calculated for every speaker's speech. The resulting embedding is a list of variable-length vectors that capture the speech properties. The number of these vectors is the same for all speakers but the length of the vectors is dependent on the length of their speech. In this study, we call a single vector a *feature vector*.

Distance between embeddings

The two embedding methods used in the experiments output embeddings in the form of a list of multiple variable-length feature vectors. Each vector within these embeddings encodes distinct acoustic information about the speaker. Despite variations in the specific information encoded by each vector, corresponding vectors from different embeddings can still be compared.

The distance between two embeddings is calculated using Dynamic Time Warping (DTW) [1], which is also known as the sequence alignment algorithm. All considered embedding methods yield variable-length feature vectors, and DTW is well-suited for this task. Moreover, speed differences that potentially propagate from the speech to the embedding are also considered in DTW. The total distance is the average of distances between corresponding feature vectors. We call this distance an *acoustic distance*.

Evaluation of the relationship between the bias and acoustics

For each speech file, the acoustic distance from the minimum group is matched with its corresponding bias, resulting in individual scatter plots for each ASR model and acoustic-based embedding. Pearson correlation coefficients are computed to determine the strength of the relation. The acoustic embedding demonstrating the highest correlation across various

ASR models is identified as the most effective method for reflecting the bias.

4 Results

Experimental setup

The data used in this study is a subset of the Jasmin dataset containing 100 speakers. Bias calculations are performed on a per-speaker basis, as the group-level calculation would result in too few data points to calculate the correlation. Acoustic embeddings are computed for the entire speech, though for practical reasons, distances between embeddings are calculated based on the first 10 feature vectors out of 1024 that comprise each embedding. The implementation is inspired by the code used in the study conducted by Bartelds [7].

Additionally, in human-machine interaction (HMI) speech recordings, the audio is stereo, with one channel capturing the speaker's speech through a microphone while the other channel records the text-to-speech prompts the speaker interacts with. Both channels are averaged into the mono format to mitigate any potential interruptions in speech flow caused by these prompts. This ensures that pauses due to text-to-speech prompts are not interpreted as hesitations from the speaker.

Results

Table 1: WER and bias for different models on Read and HMI speech types.

ASR system	Read		HMI	
	WER	**Bias**	**WER**	**bias**
NoAug	0.418	0.350	0.446	0.334
SpAug	0.400	0.322	0.541	0.349
SpSpecAug	0.384	0.305	0.477	0.299
Ws	0.406	0.307	0.496	0.319
WsFT_cgn	0.409	0.291	0.520	0.370

Table 1 summarizes the mean Word Error Rates (WERs) and biases across different ASR models for both Read and HMI speech types. Generally, biases are comparable between the two speech types, while WERs tend to be higher for HMI speech.

Table 2: Correlation between the bias and acoustic distance for different models with w2v2 and XLSR for HMI and Read speech types.

ASR system	Read		HMI	
	w2v2	**XLSR**	**w2v2**	**XLSR**
NoAug	0.594	0.550	-0.023	-0.264
SpAug	0.538	0.445	0.191	-0.069
SpSpecAug	0.553	0.500	0.362	0.030
Ws	0.537	0.471	0.029	-0.088
WsFT_cgn	0.473	0.425	-0.052	-0.042

In Table 2, the correlation values between different ASR models and acoustic embeddings (w2v2 and XLSR) are presented for Read and HMI speech types. Notably, w2v2 shows higher correlations with bias compared to XLSR for Read speech. Conversely, for HMI speech, significant correlations are generally absent across most models, with some even displaying unexpected negative correlations, suggesting randomness in the results. Only w2v2 with SpSpecAug exhibits a slight correlation above 0.3.

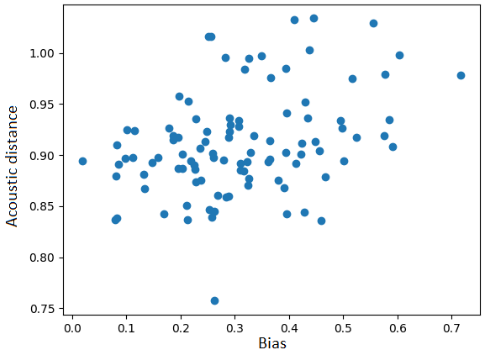

Figure 2: Scatter plot for the acoustic distance of the w2v2 embedding against the bias for the SpSpecAug model on the HMI speech.

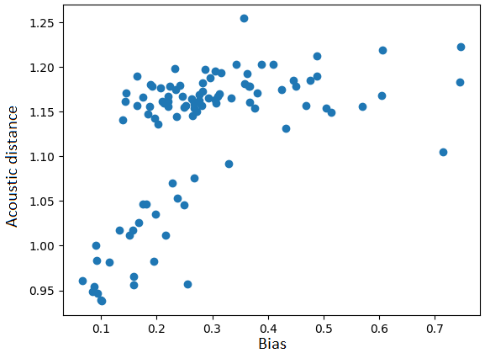

Figure 3: Scatter plot for the acoustic distance of the w2v2 embedding against the bias for the SpSpecAug model on the Read speech.

For HMI speech using the ASR model with the highest correlation (SpSpecAug and w2v2), the scatterplot in Figure 2 illustrates that the data points do not exhibit a clear correlation line. In contrast, for the Read speech (Figure 3, there is a noticeable trend where the acoustic distance increases linearly for some speakers with low bias. However, there are also instances where the distance remains similar despite variations in bias. Similar patterns are observed across all ASR and acoustic-embedding models in their respective scatterplots.

Figures 4, 5, and 6 present the scatterplots for Read speech

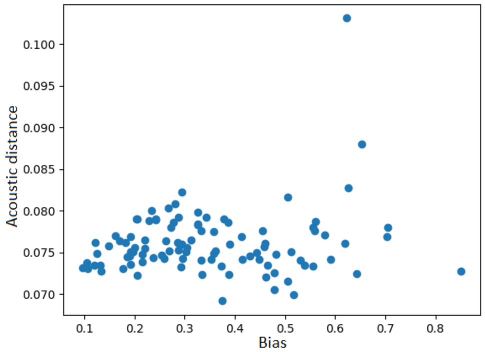

Figure 4: Scatter plot for the acoustic distance of the w2v2 embedding against the bias for the NoAug model on the Read speech. The first feature vector is used to calculate the distance.

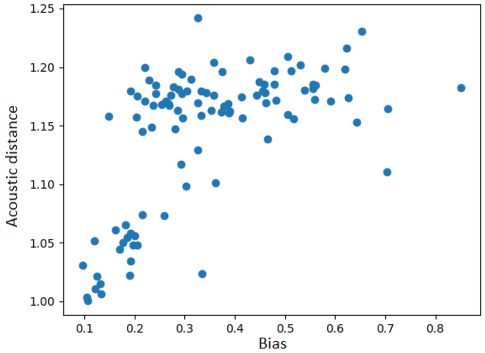

Figure 5: Scatter plot for the acoustic distance of the w2v2 embedding against the bias for the NoAug model on the Read speech. The first 10 feature vectors are used to calculate the distance.

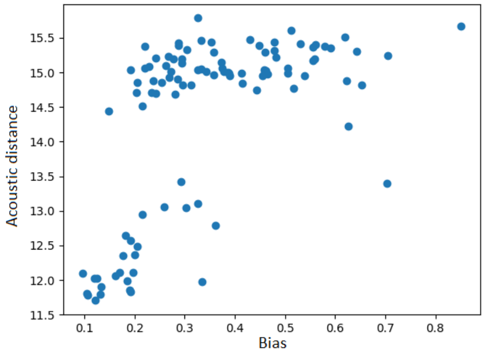

Figure 6: Scatter plot for the acoustic distance of the w2v2 embedding against the bias for the NoAug model on the Read speech. All 1024 feature vectors are used to calculate the distance.

for the NoAug model and w2v2, using 1, 10, and 1024 (all) feature vectors to calculate the acoustic distance, respectively. While using only one feature vector does not resemble the scatter plot for all the feature vectors, using ten shows resemblance, and the correlation values are similar.

Discussion

HMI speech shows little correlation between bias and any acoustic embedding model across all ASR systems. This lack of correlation might be because the recordings include both microphone input from speakers and a channel for text-to-speech prompts, which acts as noise. If the text-to-speech prompts were consistently the same in content and order, their presence might not affect the embedding distances significantly because they would align closely during the DTW alignment. However, in the HMI dataset used here, the text-to-speech prompts varied between recordings and were not scripted.

In contrast, the Read speech demonstrates a clear correlation with bias. The scripted nature of Read speech recordings, where speakers say the same words under controlled conditions, helps isolate acoustic features. This controlled setup allows for capturing distinct acoustic differences between speakers, regardless of variations in vocabulary or other factors.

For the Read speech, the correlation is largest for the NoAug model (see Table 2). This could be because this model exhibits higher bias than other models, which could mean that the bias is more dependent on the acoustics of the speakers than in other models.

The number of feature vectors used also affects the accuracy of acoustic distance measurements. While using the entire acoustic embedding provides the most accurate results, even using a small subset of feature vectors—less than one percent in the case of the w2v2 model—can effectively capture speaker acoustics (see Figures 5 and 6).

5 Responsible Research

To ensure reproducibility, detailed information is provided so that others can verify and reproduce the results. The code used to produce the results is made publicly available, along with the specific library versions used. Although one of the datasets used in the experiments is restricted to research purposes, its structure is thoroughly described. This allows for the possibility of reproducing the results with a similarly structured dataset.

We believe the nature of this research is ethical. Bias in software is important both socially and in terms of software quality. Socially, equal treatment and rights for all people are fundamental ethical principles. From a software quality perspective, it is important that software performs consistently across diverse user groups, as overfitting to specific groups is generally undesirable.

6 Conclusion and Future Work

The bias of the ASR system moderately correlates with the acoustics of the speaker. The study shows that wav2vec 2.0 acoustic embedding, in combination with the DTW, has the potential to quantify the acoustics of the speaker and explain the bias associated with it. Moreover, the findings indicate that a small subset of the wav2vec 2.0 embedding represents the acoustics of the speaker well relative to the whole embedding.

Despite the study considering multiple ASR systems and acoustic embeddings, the fact that these are all design choices cannot be overlooked. While the ASR systems used in the study are considered state-of-the-art, what defines a good acoustic embedding or bias metric is not set in stone. In particular, the acoustic embeddings in this study were based on neural network approaches, but perhaps, to get more insight into what acoustic features play the most important role in causing could be identified by looking at isolated speech properties such as energy, pith, or the length of vowels.

Although a statistically significant correlation was found between bias and acoustics, it could be that the relation is not linear. In fact, after some threshold of bias value, the acoustic distance remained within a fixed range. This hints at a nonlinear interrelation between bias and acoustics and should be investigated with statistical metrics other than correlation.

To better understand which acoustic features and pronunciations of specific words are related to bias, it would be beneficial to isolate the speech into fragments with single sentences or words. Identifying these features would help pinpoint the potential causes of bias.

It would also be worth trying more acoustic embeddings if they relate to bias. Perhaps acoustic embeddings that were not considered in this study due to, for example, poor correlation to the manual labels to native likeness would correlate much with the bias.

Another area for future work is investigating the impact of noise on the quality of acoustic embeddings. The disparity in results for read and HMI speech suggests that the acoustic embeddings used in this study do not handle noise well.

In addition to acoustic features, the vocabulary of the speaker contributes to their profile. Investigating the relationship between a speaker's vocabulary and bias could provide further insights.

Acoustic embeddings could be used not only to estimate bias but also to predict the accuracy of an ASR system. While bias is an important metric, accuracy is often the preferred measure of ASR system performance.

Finally, this study focused only on the Dutch language. It is unknown whether the results would be similar for other languages. Conducting similar studies in different languages would reveal how the variability in the acoustics of speakers affects the relation with the bias in ASR systems across different languages.

References

[1] *Dynamic Time Warping*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[2] Mohammad Abushariah and Majdi Sawalha. 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. 01 2013.

[3] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2012.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[5] Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137, May 2022. Publisher: Elsevier BV.

[6] Martijn Bartelds, Caitlin Richter, Mark Liberman, and Martijn Wieling. A New Acoustic-Based Pronunciation Distance Measure. *Frontiers in Artificial Intelligence*, 3, May 2020. Publisher: Frontiers Media SA.

[7] Martijn Bartelds and Martijn Wieling. Quantifying Language Variation Acoustically with Few Resources. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3735–3741, Seattle, United States, July 2022. Association for Computational Linguistics.

[8] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[9] Catia Cucchiarini, Hugo Van hamme, Olga van Herwijnen, and Felix 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 Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

[10] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

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

[12] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying bias in automatic

speech recognition. *arXiv preprint arXiv:2103.15122*, 2021.

[13] Sofia Kanwal, Sohail Asghar, and Hazrat Ali. Feature selection enhancement and feature space visualization for speech-based emotion recognition. *PeerJ Computer Science*, 8:e1091, 2022.

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

[15] Laura L. Koenig, Jorge C. Lucero, and Elizabeth Perlman. Speech production variability in fricatives of children and adults: Results of functional data analysis. *The Journal of the Acoustical Society of America*, 124(5):3158–3170, 11 2008.

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[17] T. Patel, W. Hutiri, A. Ding, and O. Scharenborg. How to evaluate automatic speech recognition: Comparing different performance and bias measures. unpublished.

[18] Maureen De Seyssel, Guillaume Wisniewski, Emmanuel Dupoux, and Bogdan Ludusan. Investigating the usefulness of i-vectors for automatic language characterization. In *Speech Prosody 2022 - 11th International Conference on Speech Prosody*, Lisbonne, Portugal, May 2022.

[19] Manjul Tiwari. Speech acoustics: How much science? *Journal of Natural Science, Biology and Medicine*, 3(1):24, 2012.