Gender and Dialect Bias in YouTube's Spanish Captioning System

Iris Dania Jimenez

Supervisor: Christoph Kern

Advanced Methods in Social Statistics and Social Data Science

Summer Semester 2024

Ludwig-Maximilians-Universität München

Abstract

Spanish is the official language of twenty-one countries and is spoken by over 441 million people. Naturally, there are many variations in how Spanish is spoken across these countries. However, YouTube offers only one option for automatically generating captions in Spanish. This raises the question: could this captioning system be biased against certain Spanish dialects? This study examines the potential biases in YouTube's automatic captioning system by analyzing its performance across various Spanish dialects. By comparing the performance of captions for female and male speakers from different regions, we aim to identify any systematic disadvantage faced by certain groups. Code available here.

Contents

1. Introduction

Spanish is the official language in 21 countries and is spoken by over 441 million people globally (Moreno-Fernández & Otero, 2007). Known as Español, it holds a significant presence across various continents, including Europe, where it is the dominant language in Spain, and the Americas, where it is the primary language in most of Latin America. In the United States, Spanish is widely spoken as both a first and second language, particularly in states with large Hispanic populations such as California, Texas, and Florida. The language's reach extends even to Africa, where it is the official language of Equatorial Guinea, and to Antarctica, where Chilean and Argentinian research stations maintain Spanish as a working language. Notably, Spanish is one of the six official languages of the United Nations (United Nations, 2024), highlighting its importance in international diplomacy and global affairs. It is the second most spoken language by native speakers, after Mandarin Chinese, and ranks as the third most used language on the internet (Internet World Stats, 2024). Furthermore, Spanish is the fourth most spoken language worldwide, following English, Mandarin Chinese, and Hindustani, and it is the most widely spoken Romance language(Eberhard et al., 2022). The Spanish language exhibits substantial regional variations, with major dialects such as Castilian, Mexican, Caribbean, Andean, Chilean, Paraguayan and Rioplatense Spanish(Hualde, 2005). These dialects differ in vocabulary, pronunciation, and grammar, posing challenges for standardization in media and technology. Recognizing and accommodating these differences is crucial for technology platforms, particularly those that rely on speech recognition and text generation, to ensure accuracy and accessibility for all Spanish speakers. The challenge lies in developing systems that can accurately interpret and transcribe speech from speakers of different dialects without losing the nuances that make each dialect unique. This is particularly important for educational tools, automated customer service systems, and any application where clear communication is essential.

YouTube, as a leading global platform for content creation and consumption, serves billions of users worldwide and has a unique role in facilitating language learning and cultural exchange. With over two billion logged-in monthly users (Insight, 2022), YouTube has become an essential resource for a wide array of activities. These range from educational tutorials, where users can learn new skills or languages, to entertainment and news, where users stay informed and engaged with global events. The platform's vast repository of content includes videos in countless languages, making it a versatile tool for learning and engagement for billion of users worldwide

The platform's accessibility features, such as automatic captions, are vital for ensuring that its content is inclusive and available to all users, regardless of their language proficiency or hearing ability. Automatic captions help non-native speakers understand content in different languages, support individuals with hearing impairments, and enhance the overall accessibility of videos. These features are particularly important for educational and informational content, where accurate and accessible captions can significantly enhance comprehension and learning outcomes. For instance, a student learning Spanish might rely on YouTube captions to better understand the nuances of the language, while a professional might use captions to follow along with a tutorial in a language they are less familiar with.

Despite the extensive regional variations in the Spanish language, YouTube currently provides only a single option for generating Spanish captions(Youtube, 2024). This approach does not account for the significant differences in vocabulary, pronunciation, and grammar across various Spanish dialects. As a result, the captions may not accurately reflect the spoken content for speakers of different dialects, leading to potential misunderstandings and reduced accessibility. For example, a captioning system trained primarily on Castilian Spanish might struggle to accurately transcribe Caribbean Spanish, which could result in captions that are confusing or incorrect for viewers from that region.

This study aims to evaluate the performance of YouTube's captioning system across different Spanish dialects for male and female speakers and identify any existing biases. By systematically analyzing how the system handles various dialects and genders, this research will shed light on whether the current captioning system adequately serves the diverse Spanish-speaking population or if there are gaps that need to be addressed. Understanding these biases is crucial for improving the accessibility and accuracy of automated captioning, ensuring that all users, regardless of their linguistic background, can fully benefit from the platform's offerings.

2. Related Work

2.1 Spanish Dialects

Spanish is a language rich in dialectal diversity, with significant variations across different regions. The major widely recognized dialects include Castilian, Mexican, Central American, Caribbean, Paraguayan, Chilean, Rioplatense and Andean Spanish, with each dialect exhibiting regional phonetic, lexical, and grammatical characteristics (Hualde, 2005). These dialectal differences are not just linguistic pecularities but are deeply rooted in the historical colonization, migration patterns, and interactions with indigenous languages and other European influences over the centuries. A visual geographical distribution of these dialects is shown in Figure 1.

Castilian Spanish, which is predominant in Spain, is often considered the standard form of the language in educational and media contexts. It is characterized by its use of the /θ/ sound for the letters «c» and «z» before «i» and «e», a feature known as «distinción». This phonological feature sets Castilian apart from many other Spanish dialects and has become a symbol of Spanish identity within Spain. In contrast, Latin American Spanish, which includes a broad array of regional variations, generally does not distinguish between the /s/ and /θ/ sounds, a phenomenon known as «seseo». This lack of distinction is one of the most prominent phonological differences between European and Latin American Spanish, highlighting how geographic separation and colonial history have led to divergent linguistic evolutions.

Mexican Spanish, mainly spoken in Mexico and the southern regions of the United States, is particularly interesting due to its incorporation of numerous indigenous terms, a reflection of the country's rich pre-Columbian history(Hualde, 2005). Additionally, it features distinctive diminutives and the consonant cluster /tl/, inherited from the indigenous pre-Columbian language that profoundly influenced Mexican Spanish. This cluster is particularly challenging for speakers of other Spanish dialects to pronounce, underlining the unique phonetic inventory that has developed in Mexico over centuries (Hualde, 2005). The Central American Spanish is spoken in Guatemala, El Salvador, Honduras, Nicaragua and Costa Rica. In the central nations of El Salvador, Honduras, and Nicaragua, the /s/ sound at the end of a syllable or before a consonant is often pronounced as [h], though this is less common in formal speech such as TV broadcasts(Lipski, 2008). Caribbean Spanish is spoken in Cuba, Dominican Republic, Puerto Rico, Panama and the coasts of Venezuela and Colombia. It closely resembles the Spanish spoken in the Canary Islands and, to a lesser extent, the Spanish of western Andalusia. It is noted for its rapid speech and the aspiration or omission of the /s/ sound at the end of syllables(Lipski, 2008). In Paraguay, Spanish coexists with Guarani, an indigenous language that has official status alongside Spanish. Paraguayan Spanish, also spoken in the lowlands of Eastern Bolivia (Hualde, 2005), exhibits distinctive features reminiscent of the Spanish formerly spoken in northern Spain. This is due to the significant number of early Spanish colonizers originating from the Basque Country. The Guarani language has greatly influenced Paraguayan Spanish, affecting both its vocabulary and grammar, leading to a unique linguistic blend that reflects the country's dual linguistic heritage (Hualde, 2005). Chilean Spanish is primarily spoken in Chile and the neighbouring areas. The Royal Spanish Academy recognizes 2,214 words and idioms exclusively or mainly produced in Chilean Spanish, in addition to many still unrecognized slang expressions('Nuevo diccionario ejemplificado de chilenismos y de otros usos diferenciales del español de Chile. Tomos I, II y III', 2020), Chilean Spanish is also notoriously known among Spanish native speakers to be one of the most different dialects(Alemany, 2021). Rioplatense Spanish, spoken in Argentina and Uruguay, is distinctive for its intonation, which often resembles the Neapolitan language of Southern Italy. This feature is a legacy of the massive Italian immigration to Argentina and Uruguay in the late 19th and early 20th centuries (Zenkovich, 2018). The use of the pronouns «vos» instead of «tú» for informal address is another characteristic feature of Rioplatense Spanish, differentiating it from other Spanish dialects. Andean Spanish is a dialect spoken in the central Andes, stretching from southern Colombia to northern Chile and northwestern Argentina, and encompassing Ecuador, Peru, and Bolivia. This dialect, while similar to other forms of Spanish, is heavily influenced by indigenous languages such as Quechua and Aymara(Hualde, 2005).

These dialectal distinctions present challenges for standardizing the language and for technological applications like automatic speech recognition, which must accommodate this linguistic diversity to perform accurately and inclusively. Understanding and accounting for these regional variations is crucial for developing systems that can accurately transcribe and interpret spoken Spanish across different dialects.

2.2 Automatic Speech Recognition Systems

The development of Automatic Speech Recognition (ASR) systems has a rich history dating back to the 1950s. Initially, these systems were limited to recognizing isolated digits and small vocabularies (Juang & Rabiner, 2005). One of the earliest notable examples was IBM's «Shoebox» from the 1960s, which could understand and respond to a small set of spoken commands. However, these early systems were far from perfect, often requiring speakers to enunciate clearly and pause between words to achieve any level of accuracy. These limitations highlighted the complexity of human speech and the challenges in developing systems that could mimic the natural language processing capabilities of the human brain.

Significant advancements in ASR technology occurred in the 1970s and 1980s with the introduction of Hidden Markov Models (HMMs). HMMs allowed ASR systems to model the temporal variations in speech, significantly improving accuracy by enabling the system to predict the probability of a sequence of phonemes rather than relying on static, isolated sounds (Jelinek, 1997). This advancement marked a pivotal shift from simple pattern recognition towards more sophisticated statistical modeling, paving the way for

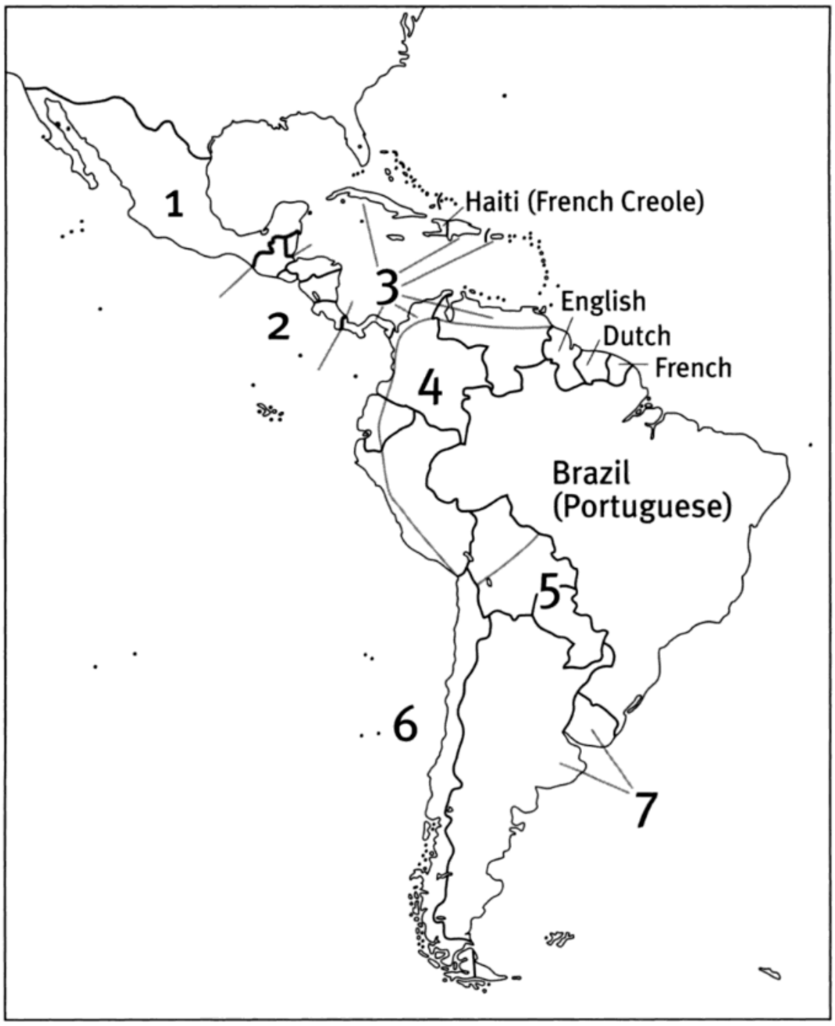

Figure 1: Table describing the Latin American dialects from (Hualde, 2005). (1) Mexican; (2) Central America; (3) Caribbean; (4) Andean; (5) Paraguayan; (6) Chilean; (7) Rioplatense

more complex and capable ASR systems.

By the 1990s, the field of ASR technology experienced further improvements with the integration of statistical language models and the rise of large-vocabulary continuous speech recognition (LVCSR) systems (Young, 1996). These systems were capable of understanding continuous speech, where words are spoken in a flow rather than in isolation, which is closer to how people naturally speak. This development was made possible by the increasing computational power available at the time, which allowed for the processing of larger datasets and more complex algorithms. Additionally, the availability of large-scale annotated corpora enabled the training of more robust models that could handle the variability inherent in human speech, such as differences in accent, intonation, and speaking style.

In recent years, ASR technology has seen remarkable advancements, particularly in languages like English and Chinese. These improvements have been driven by extensive research, the availability of vast datasets, and the application of sophisticated machine learning algorithms, particularly deep learning techniques. For example, Google's ASR system has achieved high accuracy rates in English, largely due to the availability of large, diverse datasets and continuous technological enhancements (Saon et al., 2016). The system's success can be attributed to its ability to learn from a vast amount of data, encompassing a wide range of accents, speech patterns, and contextual uses of language. Similarly, Chinese ASR systems have benefited from targeted research efforts and the integration of tonal and phonetic elements unique to the language, leading to robust and reliable performance (Singh & Kadyan, 2020). The success of these systems in handling languages with complex tonal and phonetic structures showcases the versatility and adaptability of modern ASR technologies.

Despite these advancements, research on ASR bias, particularly in relation to Spanish and its regional dialects, remains limited. The Spanish language is spoken by over 441 million people across 21 countries, each with its own regional dialects and variations in pronunciation, vocabulary, and syntax. This diversity poses a significant challenge for ASR systems, which must be able to accurately recognize and transcribe speech from speakers with different linguistic backgrounds. In recent years, studies have demonstrated that the performance of ASR systems often declines when tested on dialectal variations of a language that were not included in the training data (Elfeky et al., 2018; Chan et al., 2022). This decline in performance highlights the importance of training ASR systems on diverse datasets that include a wide range of dialects and speaking styles. However, the challenge remains that training such systems on a multitude of dialects can dilute their effectiveness for any single dialect, as models trained on multiple dialects at once tend to be less effective than those specifically trained for individual dialects (Parsons et al., 2023; Chan et al., 2022).

The issue of dialect bias in ASR systems has been explored in several languages beyond Spanish. For instance, different studies have been conducted on dialect bias for Arabic (Droua-Hamdani et al., 2012; Sawalha & Shariah, 2013) and English (Wheatley & Picone, 1991; Tatman, 2017; Markl, 2022) speakers. These studies have found that ASR systems often perform better on the dialects or accents that are more prominently represented in the training data, leading to disparities in performance that can disadvantage speakers of less common or non-standard dialects. However, the literature still lacks a comprehensive study on Spanish dialect bias across different ASR systems. This gap is significant given the widespread use of Spanish globally and the increasing reliance on ASR technology in everyday applications such as virtual assistants, automated transcription services, and language learning tools.

Nevertheless, considerable work has been done to improve current ASR systems or create new ones for the Spanish language. Efforts have included implementing a single multidialectal model to accommodate the diverse Spanish dialects spoken across Europe and Latin America (Caballero et al., 2009). Additionally, researchers have developed regionalized models for Spanish language variations based on Twitter data, which offers a rich and diverse source of linguistic information (Tellez et al., 2023). Furthermore, there has been work on creating automatic dialect recognizer systems specifically for Mexican (Hernández-Mena et al., 2017), Cuban, and Peruvian Spanish (Zissman et al., 1996). These recognizer systems aim to improve the accuracy of ASR systems by tailoring them to specific dialects, thereby reducing the errors that arise from dialectal variation. Other efforts have focused on improving the resilience of ASR models against different native Spanish accents (Chitkara et al., 2022) and performing punctuation restoration for speech-to-text ASR systems (Zhu et al., 2022). These advancements reflect a growing recognition of the need to address linguistic diversity in ASR systems and the ongoing efforts to make these systems more inclusive and accurate for all users.

So far, this discussion has primarily focused on the biases that ASR systems can exhibit based on different dialects. However, in the literature, there has been a significant amount of work focusing on the gender of the speakers as well. Male and female voices have different acoustic characteristics, such as pitch, tone, and speech patterns (Gelfer & Mikos, 2006). These differences have been shown to affect the performance of ASR systems, with studies demonstrating bias against female speakers (Garnerin et al., 2019) and, in some cases, bias against male speakers (Sawalha & Shariah, 2013; Feng et al., 2021; Adda-Decker & Lamel, 2005). These studies reveal that while ASR systems are improving, they are not yet fully equitable across all demographics. It's important to note that none of these studies included non-binary or transgender speakers, highlighting a gap in the research that needs to be addressed to ensure ASR systems are inclusive of all users.

Furthermore, ASR systems have also been shown to work better for younger speakers compared to older speakers (Sawalha & Shariah, 2013). This age-related bias likely stems from the fact that most training data for ASR systems comes from younger adults, leading to a system that is better tuned to the speech patterns of this demographic. Additionally, studies have shown that ASR systems exhibit a biased performance when comparing English white speakers and English African American speakers, with worse performances for the latter group (Koenecke et al., 2020). This racial bias in ASR systems has serious implications for their use in real-world applications, particularly in areas such as law enforcement and customer service, where accurate speech recognition is crucial. Furthermore, speech disabilities can impact the performance of ASR systems, with research showing that individuals with speech impairments often experience higher error rates when using these technologies (Moro-Velazquez et al., 2019; Halpern et al., 2020). These findings underscore the need for more inclusive training data that represents a wider range of speech patterns and the development of ASR systems that are robust to variations in speech.

ASR systems have thus shown disparities in performance across various biases, including age, gender, dialects, and speech disabilities. Extensive work has been done to study these biases, but more needs to be done to address them comprehensively. The most similar work in literature to this study are (Tatman, 2017), where the quality of YouTube's automatically generated captions was tested for gender and different English dialects. Bias was detected against women and speakers from Scotland, indicating that even widely used systems like YouTube's ASR are not immune to these issues. Another relevant study is (Elfeky et al., 2018), where the Google Assistant Voice system was tested for five different Spanish dialects (US, Spain, Mexico, Argentina, and Latin America). The study found that the system's performance varied across these dialects, further highlighting the need for more targeted improvements in ASR systems. To the best of my knowledge, there has not yet been a study that analyzes YouTube's captioning system performance for both gender and Spanish dialects, making this study a contribution to the field.

3. Data

For this work, I used two datasets: the Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech dataset (Guevara-Rukoz et al., 2020) for Latin American dialects, and the TEDx Spanish Corpus (Hernandez-Mena, 2019) for the Spain dialect. The Latin American dataset covers six countries: Argentina, Chile, Colombia, Peru, Puerto Rico, and Venezuela. This corpus consists of crowdsourced recordings from both male and female speakers, along with their corresponding orthographic transcriptions. Each dataset is divided into two subsets, one for female speakers and the other for male speakers. All recorded volunteers were native speakers of their respective dialects. Recordings for the Argentinian, Chilean, Colombian, and Peruvian dialects were conducted in their native regions, while recordings for the Puerto Rican and Venezuelan dialects took place in New York, San Francisco, and London. The original recording script, designed for Mexican Spanish, was adapted for the different dialects by shortening phrases and removing references specific to Mexican Spanish. Additional sentences were generated using templates to increase variety. Although only a small portion of the script was specifically adapted by native speakers for each dialect, speakers were allowed to improvise to ensure a natural representation of their dialects. Any mismatches between transcriptions and audio were corrected during quality control. The script included approximately 30 «canonical» sentences across all dialects to capture phonological contrasts. Dialect-specific pronunciation lists were expanded to cover more words, with manual edits to correctly capture the pronunciation of loanwords. In total, the recordings for the Latin American dataset comprise 19 hours of female speakers and 18 hours of male speakers, totaling 37 hours of audio with 176 unique speakers. For Puerto Rico, only female speakers were recorded. A more detailed description of the dataset can be found in Figure 2. The audio was recorded as 48 kHz single-channel and is provided in 16 bit linear PCM RIFF format. The TEDx Spanish Corpus is a 24-hour, gender-imbalanced dataset featuring spontaneous speeches from various TEDx event presenters. This dataset contains 11243 audio files from 142 different speakers, 102 of whom are male and 40 female. Each audio file is approximately 3 to 10 seconds long. Audios and transcriptions are provided in lowercase without punctuation, and all the transcriptions were done by native Spanish speakers. The audio files are distributed in Windows WAVE 16 kHz @ 16-bit mono format.

For each dataset, every audio recording was assigned a unique identifier to distinguish between multiple recordings from the same speaker. This careful organization was necessary because each speaker contributed more than one audio sample. In total, combining both the datasets from Spain and Latin America, a comprehensive collection of 69 hours of audio was gathered.

Dialect	Code	Locations	ISLRN	Gender	Name	Lines	Words Total	Words Unique	Duration (hours)	Speakers
Argentinian	AR	Buenos Aires	395-001-133-368-2	F	arf	3,921	35,360	4,107	5.61	31
				M	arm	1,818	16,914	3,343	2.42	13
Chilean	CL	Santiago	048-218-632-043-6	F	clf	1,738	16,591	3,279	2.84	13
				M	clm	2,636	25,168	4,171	4.31	18
Colombian	CO	Bogota	169-985-498-793-0	F	cof	2,369	22,228	4,460	3.74	16
				M	com	2,534	23,957	4,459	3.84	17
Peruvian	PE	Lima	923-742-092-167-6	F	pef	2,529	23,806	4,278	4.35	18
				M	pem	2,918	27,547	4,268	4.87	20
Puerto Rican	PR	US	721-732-548-994-0	F	prf	617	6,092	1,738	1.00	5
				M	–	–	–	–	–	–
Venezuelan	VE	US and UK	697-927-390-879-1	F	vef	1,603	15,182	3,419	2.41	11
				M	vem	1,754	16,613	3,612	2.40	12
Total:						24,437	229,458	5,783	37.79	174

Figure 2: Table describing the Latin American dataset from (Guevara-Rukoz et al., 2020)

4. Methodology

The goal of this study is to evaluate the performance of YouTube's captioning system for male and female speakers across different Spanish dialects. To achieve this, the first step was determining the optimal format for uploading the audio files to YouTube. The audio files, each averaging five seconds in length, were too numerous to upload individually, as doing so manually would have been highly impractical. Given the limitations of manually uploading, I utilized the YouTube API, which allows video uploads via the command line. However, the API imposes a daily upload limit of six videos, making it impractical to upload each five-second clip individually within a reasonable timeframe, given that in total I had 69 hours of audios. So far, I have utilized the term «audios», but of course, YouTube only accepts video formats, not audio files. Therefore, before uploading the final results to YouTube, I had to convert the audios to videos. I did this by simply using a black image as the background for all the audios.

My initial strategy was to aggregate the audio files by gender and country, creating a single long audio file for each group, for example, combining all the audio clips of female speakers from Puerto Rico into one large file. The main challenge with this approach was that, by combining the audio files into a single video, I needed to maintain information about which audio clip was at which timestamp. This was crucial for mapping the generated captions back to the corresponding ground truth data. To address this, I created a mapping file that tracked each audio's ID and its timestamp within the video, ensuring that I could later compare the generated captions with the ground truth data via the speaker's ID. This approach reduced the number of videos to thirteen, which could theoretically be uploaded in three days via the YouTube API.

However, the resulting videos were approximately five hours long each. Uploading such lengthy videos proved to be computationally expensive and ultimately unfeasible due to the limitations of YouTube's API in handling and processing long videos. To overcome this challenge, I revised my approach and opted to limit the audio files to thirty minutes in length per video. While this approach does not fully utilize all the available data, it proved to be the most feasible solution given the computational constraints I faced.

The next step, after uploading the videos to YouTube, was to retrieve the captions. I accomplished this by using the YouTube API to download the captions while preserving the timestamp information, which allowed me to match the captions back to the corresponding audio segments. However, I encountered a challenge with this approach: the captions were often out of sync with the audio, a common issue reported by users of YouTube's automatic captioning system. To mitigate this problem, I added a five-second delay between each audio segment and then re-uploaded the videos to YouTube. While this adjustment did not completely resolve the issue, it significantly reduced the impact of the synchronization problem.

This methodology was carefully designed to manage the challenges of working with You-

Tube's captioning system and the limitations of its API. Despite the challenges involved, it allowed for a thorough evaluation of the system's performance across various Spanish dialects and genders. This approach ensured that the analysis was as detailed and accurate as possible, given the constraints, and provided valuable insights into how well YouTube's ASR technology handles different types of Spanish speakers and dialects.

5. Results

To evaluate the accuracy of the generated captions in comparison to the annotated ground truth data, I employed the Word-Error-Rate (WER) metric. WER is a well-established and widely used metric in the field of Automatic Speech Recognition to assess the performance of systems handling large vocabularies. It provides a straightforward measure of how accurately the ASR system transcribes spoken words into text. The WER is calculated by comparing the word sequence generated by the ASR system against a reference transcription, and counting the number of errors, which include substitutions (S), insertions (I), and deletions (D). These errors are then summed and normalized by the total number of words (N) in the reference transcription. The formula for WER is expressed as follows (Ali & Renals, 2018):

$$WER = \frac{I + D + S}{N} \times 100 \tag{1}$$

A lower WER indicates better accuracy in recognizing speech. For instance, a WER of 20% implies that the transcription is 80% accurate. WER can be calculated as either a case-sensitive or case-insensitive metric. Given that case-sensitive WER is most commonly used in the literature, I have chosen to adhere to this approach to maintain consistency with existing research.

In the following sections, I will present and analyze the results obtained from the WER metric across various countries, genders, and dialectal groups.

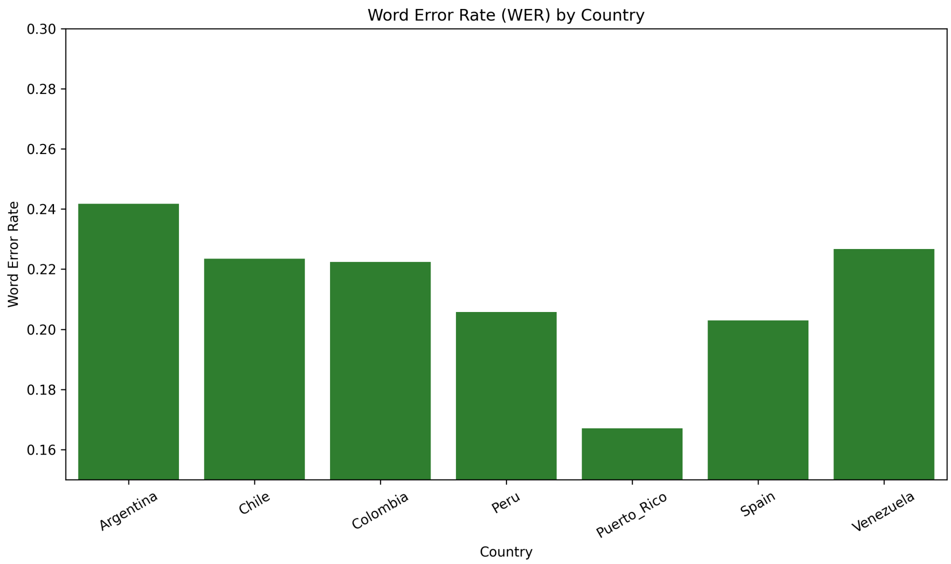

Figure 3: WER for each country

The bar plot in Figure 3 illustrates the WER obtained for each country included in the study. Youtube's ASR achieved the best performance for speakers from Puerto Rico with a WER of 16%, indicating that 84% of the generated captions accurately matched the ground truth annotations, suggesting that the ASR system has a relatively high level of accuracy when transcribing the Puerto Rican dialect, which is part of the Caribbean dialect. Following Puerto Rican dialect, Youtube's ASR exhibited strong performance on both Castillian and Peruan dialects, each with a WER of 20%. For Chilean and Colombian speakers, the generated captions were accurate in 78% of cases, corresponding to a WER of 22%. The worst performance was observed for Argentinian speakers, who experienced a WER of 24%, meaning that only 76% of the captions matched the reference transcriptions.

These results are somewhat surprising. Based on the discussion in Section 2.1, it was expected that Chilean speakers would experience the poorest ASR performance, given that the Chilean dialect is often considered the most challenging for other native Spanish speakers to understand (Alemany, 2021). The Chilean dialect is known for its unique phonetic features, rapid speech, and use of colloquial expressions that are less common in other Spanish-speaking regions. Despite these characteristics, the ASR system performed relatively well for Chilean speakers, which could suggest that the system has been trained on a diverse dataset that includes representations of this dialect.

Equally unexpected is the excellent performance for Puerto Rican speakers, particularly because Caribbean Spanish, including the Puerto Rican dialect, is generally spoken at a faster pace compared to other dialects. A possible explanation for this result could be that, since YouTube is an American company and Puerto Rico is an official territory of the United States, the training data may have included a larger representation of Puerto Rican speakers compared to those from other dialects. This greater representation could have enabled the ASR system to more accurately capture the nuances of Puerto Rican Spanish.

It's important to highlight the considerable difference in performance between the best (Puerto Rico, 16% WER) and worst (Argentina, 24% WER) results, with an 8% gap in WER, which is quite significant. This performance gap indicates that while the Youtube's ASR system may be generally effective, there are clear disparities in how well it handles different Spanish dialects. Understanding the root causes of these disparities could be a critical area for future research, as it may point to specific phonetic or lexical features that the ASR system struggles with.

Next, I analyzed the WER for male and female speakers to determine if there were any notable gender-based differences in performance.

As shown in Figure 4, the performance difference between male and female speakers is minimal, with female speakers achieving a WER of 20% and male speakers at 21%, resulting in a mere 1% difference. This small gap is still consistent with existing literature, which generally indicates that ASR systems tend to perform slightly better for female speakers compared to male speakers (Sawalha & Shariah, 2013; Feng et al., 2021; Adda-Decker & Lamel, 2005). The reasons for this difference could be multifaceted, potentially involving differences in pitch, speech rate, and articulation patterns between genders. However, the overall small difference suggests that the system is relatively balanced in its

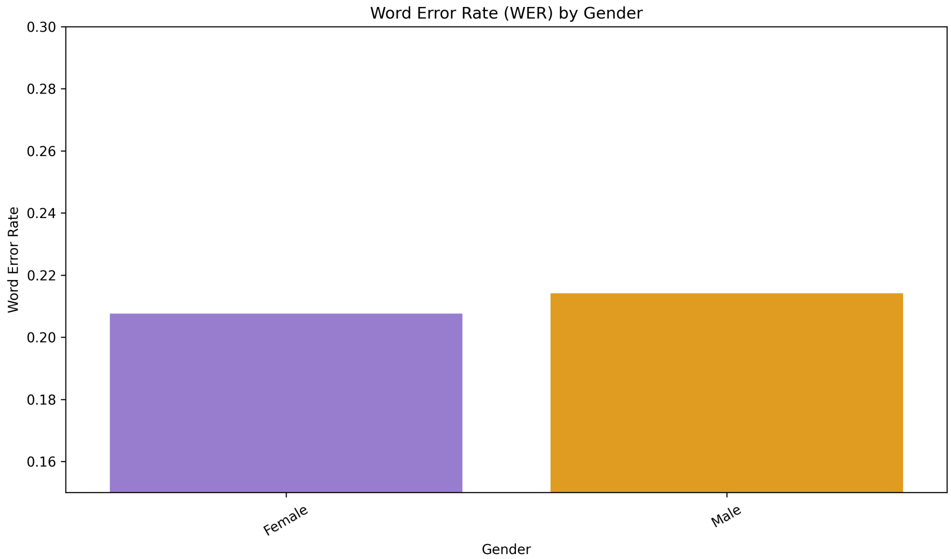

Figure 4: WER by gender across all countries

treatment of male and female voices.

Next, I examined the performance of Youtube's ASR system for each country, further stratified by gender. This analysis helps to uncover whether the observed gender differences are consistent across different dialects or if they vary significantly between regions.

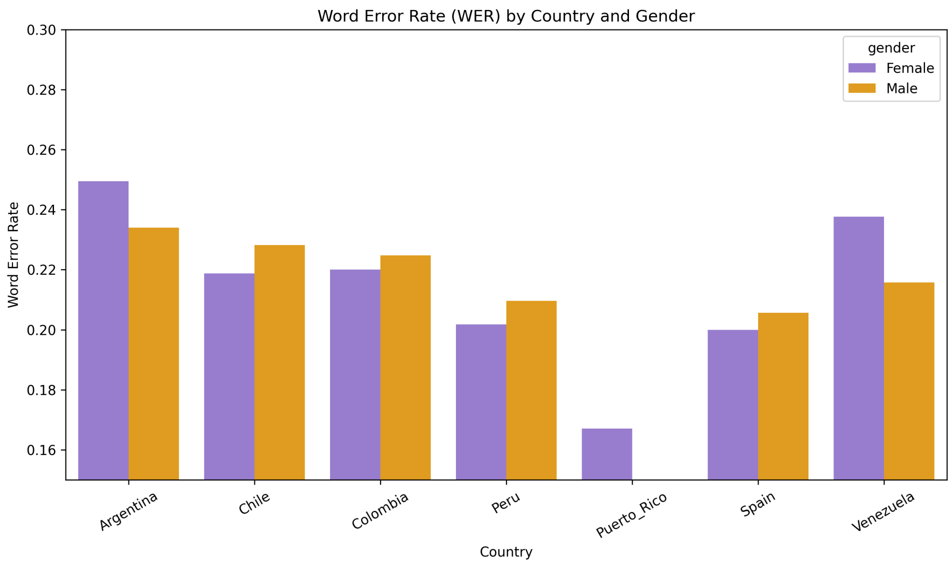

Figure 5: WER by country and gender

As shown in Figure 5, it's important to highlight that no data was available for male Puerto Rican speakers, which limits our ability to draw conclusions about gender-based performance for this dialect. However, in the cases of Chile, Colombia, Peru, and Spain, male speakers exhibited higher Word Error Rates compared to female speakers within the same dialects. This suggests that YouTube's ASR system may be slightly more attuned to the speech patterns of female speakers in these countries, although the difference between male and female WER is relatively small.

Interestingly, even though there was more training data available for male speakers in these regions, a factor that would usually be expected to result in lower WER for males, the opposite trend was observed. This indicates that other factors may be influencing the system's performance.

In contrast, in Venezuela and Argentina, male speakers achieved a lower WER, indicating better performance in the generated captions compared to their female counterparts. The performance gap between genders in these countries is more pronounced than in the other regions, suggesting that regional differences in gender-influenced speech characteristics might not be effectively captured by the ASR system. Notably, in Argentina, there were 31 audio samples from female speakers compared to only 13 from male speakers. One might expect this disparity in data to result in better performance for female speakers, yet the opposite was observed, further implying that additional factors may be at play.

Upon examining the data, it becomes evident that countries with a higher number of male speakers do not show a corresponding improvement in WER for this gender, as might be expected. Conversely, countries with more female speakers often display a higher WER for females. This suggests that the quantity of gender-specific data does not directly correlate with better performance for the more represented gender, indicating that other factors, beyond just the amount of training data, may be influencing the ASR system's effectiveness.

Lastly, I would like to comment on the difference between the performance obtained for Spain and the performance for the Latin American countries (Argentina, Chile, Colombia, Peru, Puerto Rico, and Venezuela). The rationale behind this comparison is that, as we explored in Section 2.1, there is a major difference between the Spanish and Latin American dialects. Castilian Spanish, spoken in Spain, differs significantly in phonology, vocabulary, and even some aspects of grammar from the various Latin American dialects. The comparison in terms of WER is shown in Figure 6.

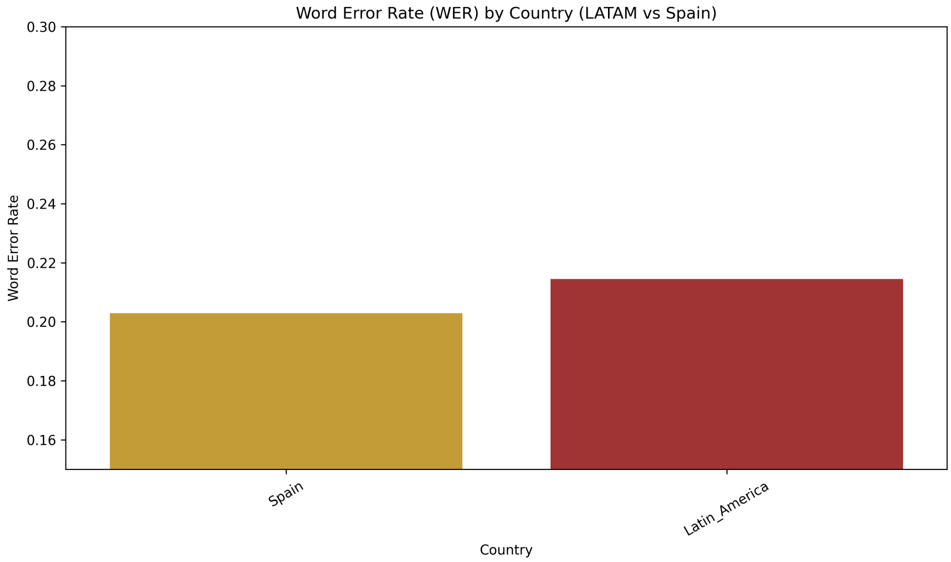

Figure 6: WER comparison between Spain and Latin America

As we can see from Figure 6, there isn't a significant difference in the performance of YouTube's captioning system between speakers from Latin America and Spain. Both groups tend to receive captions of fairly similar quality, though Latin American speakers appear to experience slightly worse performance overall. This slight discrepancy could be due to the broader range of dialectal variation within Latin America, which might pose more challenges for the Youtube's ASR system compared to the more standardized form of Castilian Spanish spoken in Spain. However, the relatively close WER values suggest that the ASR system has been trained on a diverse enough dataset to handle a wide array of Spanish dialects with reasonable accuracy.

6. Limitations and Future Directions

As observed, the best performance in terms of the quality of generated captions was achieved for Puerto Rican female speakers. This finding is particularly intriguing and raises several questions about the factors contributing to this superior performance. One possibility is that the specific phonological features of the Puerto Rican dialect are more easily recognized by YouTube's ASR system, perhaps due to better representation in the training data or alignment with the system's underlying models. If more data becomes available in the future, it would be highly valuable to explore the performance of YouTube's ASR system on male Puerto Rican speakers as well. This additional data would allow for a more comprehensive analysis, helping to determine whether the high performance observed for female speakers is consistent across genders or if there are notable variations that need to be addressed. Although the current analysis indicates minimal differences in performance between male and female speakers within the same country, it is important to validate these findings with additional data to ensure the robustness of these observations across different contexts. Furthermore, it would be valuable to explore the performance of YouTube's ASR system across various Caribbean dialects, such as Dominican or Cuban Spanish. By conducting these tests, we could determine whether the system's relatively strong performance with Puerto Rican Spanish is an isolated case or if it extends to other Caribbean dialects as well. This broader analysis would provide deeper insights into the system's adaptability and accuracy across different regional variations of Spanish, offering a more comprehensive understanding of its strengths and limitations in processing diverse dialects from the Caribbean region.

The significant difference in performance between Argentinian and Puerto Rican speakers is particularly noteworthy. This disparity suggests that there are underlying factors, whether linguistic or technical, that influence the accuracy of the Youtube's ASR system across different dialects. Understanding whether these differences stem from specific phonetic characteristics, such as the pronunciation of certain consonants and vowels, or from broader systemic issues within the ASR models, would be a valuable direction for future research.

It is also important to note that this analysis focused on only six Spanish-speaking countries, while there are many more countries where Spanish is the official language, each with its own unique dialectal variations. Expanding the scope of this study to include a wider range of dialects would offer a more comprehensive understanding of YouTube's ASR system performance and help identify areas that require improvement. In particular, including speakers from Equatorial Guinea, the only Spanish-speaking country in Africa, would provide a unique and valuable perspective on the system's ability to process lesser-studied Spanish dialects. Given Equatorial Guinea's distinct linguistic and cultural context, which differs significantly from the more commonly studied Latin American and European variants, testing the ASR system with speakers from this region would enhance our understanding of its global applicability and inclusivity.

Future research could also cluster speakers based on the seven major dialects outlined in Section 2.1—Castilian, Mexican, Central American, Caribbean, Paraguayan, Chilean, Rioplatense, and Andean Spanish—while also incorporating African Spanish speakers from Equatorial Guinea. This approach would facilitate a deeper exploration of the ASR system's performance across these diverse dialects and reveal whether there are significant accuracy differences that need to be addressed. By broadening the range of dialects studied, researchers can gain more insights into the effectiveness of current ASR models in capturing the linguistic diversity present within the global Spanish-speaking community, ultimately identifying specific areas where targeted improvements are necessary to enhance system performance and inclusivity..

Moreover, a more detailed analysis of the specific phonemes and linguistic features that challenge YouTube's captioning system would be highly valuable, similarly to what has been done in previous studies for Dutch and Chinese Mandarin(Feng et al., 2023). For example, breaking down the audio data to study the performance on specific phonemes, such as the /s/ and the /θ/, could reveal where the system struggles most. This phoneme-level analysis could identify specific sounds or combinations of sounds that are prone to errors, which could then be targeted for improvement in future versions of the ASR system. Additionally, understanding how the system handles regional variations in intonation, rhythm, and stress patterns could provide further insights into its strengths and weaknesses. This could lead to targeted improvements in ASR technology, making it more robust across different dialects and speech patterns. Such enhancements would not only improve the accuracy of captions for a wider audience but also contribute to the broader goal of making digital content more accessible and inclusive.

In addition to exploring gender and dialectal biases, it would be highly valuable to investigate how YouTube's ASR system performs across different age groups. Age-related biases are a known issue in speech recognition technology, as the acoustic characteristics of speech can vary significantly across different stages of life. By including a diverse range of age groups in future studies, from younger to older speakers, we can assess whether the system is equally effective across all age demographics or if certain age groups experience more errors.

A promising avenue for future research involves conducting longitudinal studies to monitor the performance of YouTube's ASR system over time. By systematically evaluating the system at regular intervals, researchers can observe how updates to the system or the introduction of new training data impact its accuracy and reliability across different Spanish dialects. This approach would help determine whether improvements are uniformly distributed across all dialects and speaker demographics or if certain groups continue to experience disparities in ASR quality.

In addition to longitudinal analysis, a cross-platform comparison could provide valuable insights into the relative strengths and weaknesses of YouTube's ASR system. By comparing it with other widely used platforms, such as Google Assistant Voice, Apple's Siri, or Amazon's Alexa, researchers can benchmark YouTube's performance against industry standards. This comparative analysis could highlight specific areas where YouTube's system excels or lags, providing a clear direction for targeted improvements and contributing to the development of more robust and inclusive ASR technologies across the board.

One major limitation of this work was the inability to fully exploit the available datasets due to computational constraints and limitations imposed by YouTube's API. These constraints prevented the use of a significant amount of additional audio data, which might have provided a more comprehensive analysis and potentially different insights into the system's performance. In future work, the goal would be to fully leverage all available audio data by uploading it to YouTube, thus enabling a more exhaustive evaluation of the ASR system's performance. Overcoming these limitations would allow for a more detailed and accurate assessment of how well the ASR system performs across different dialects and genders. Additionally, finding alternative methods or tools to circumvent the limitations of YouTube's API could be an area of exploration, as it would enable researchers to conduct more thorough analyses without being constrained by current technological barriers.

7. Conclusion

In conclusion, this study has provided valuable insights into the performance of YouTube's ASR system across different Spanish dialects, with a particular focus on gender-based differences. While the system demonstrated relatively strong performance for Puerto Rican female speakers, the results also highlighted significant disparities, such as the higher WER for Argentinian speakers and the unexpected lack of improved accuracy for genders with more available training data. These findings underscore the need for further research to better understand the linguistic and technical factors that influence Youtube's ASR performance. Future studies should aim to broaden the scope of analysis to include a wider range of dialects and demographics, such as different age groups and underrepresented regions like Equatorial Guinea. Additionally, exploring phoneme-level challenges and conducting cross-platform comparisons will be crucial in identifying specific areas for improvement. Overcoming the current limitations related to data utilization and API constraints will be essential for enabling more comprehensive evaluations and driving advancements in ASR technology, ultimately contributing to more inclusive and accurate speech recognition systems.

References

Adda-Decker, M., & Lamel, L. (2005). Do speech recognizers prefer female speakers? *Interspeech*. https://api.semanticscholar.org/CorpusID:8668656

Alemany, L. (2021). El español de chile: La gran olla a presión del idioma [Retrieved 1 June 2022]. *El Mundo*. https://www.elmundo.es/cultura/2021/11/30/619519fbfdddff4e208b45a6.html

Ali, A., & Renals, S. (2018, July). Word error rate estimation for speech recognition: E-WER. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 20–24). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2004

Caballero, M., Moreno, A., & Nogueiras, A. (2009). Multidialectal spanish acoustic modeling for speech recognition. *Speech Communication*, 51(3), 217–229. https://doi.org/https://doi.org/10.1016/j.specom.2008.08.003

Chan, M. P. Y., Choe, J., Li, A., Chen, Y., Gao, X., & Holliday, N. (2022). Training and typological bias in ASR performance for world Englishes. *Proc. Interspeech 2022*, 1273–1277. https://doi.org/10.21437/Interspeech.2022-10869

Chitkara, P., Riviere, M., Copet, J., Zhang, F., & Saraf, Y. (2022). Pushing the performances of asr models on english and spanish accents. https://arxiv.org/abs/2212.12048

Droua-Hamdani, G., Selouani, S. A., & Boudraa, M. (2012). Speaker-independent asr for modern standard arabic: Effect of regional accents. *International Journal of Speech Technology*, 15. https://doi.org/10.1007/s10772-012-9146-4

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2022). Summary by language size [Archived from the original on June 18, 2023. Retrieved December 2, 2023]. https://www.ethnologue.com

Elfeky, M. G., Moreno, P., & Soto, V. (2018). Multi-dialectical languages effect on speech recognition: Too much choice can hurt [1st International Conference on Natural Language and Speech Processing]. *Procedia Computer Science*, 128, 1–8. https://doi.org/https://doi.org/10.1016/j.procs.2018.03.001

Feng, S., Halpern, B., Kudina, O., & Scharenborg, O. (2023). Towards inclusive automatic speech recognition. *Computer Speech and Language*, 84. https://doi.org/10.1016/j.csl.2023.101567

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. https://arxiv.org/abs/2103.15122

Garnerin, M., Rossato, S., & Besacier, L. (2019). Gender representation in french broadcast corpora and its impact on asr performance. https://arxiv.org/abs/1908.08717

Gelfer, M., & Mikos, V. (2006). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of voice : official journal of the Voice Foundation*, 19, 544–54. https://doi.org/10.1016/j.jvoice.2004.10.006

Guevara-Rukoz, A., Demirsahin, I., He, F., Chu, S.-H. C., Sarin, S., Pipatsrisawat, K., Gutkin, A., Butryna, A., & Kjartansson, O. (2020, May). Crowdsourcing Latin American Spanish for low-resource text-to-speech. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 6504–6513). European Language Resources Association. https://aclanthology.org/2020.lrec-1.801

Halpern, B. M., van Son, R., Brekel, M. v. d., & Scharenborg, O. (2020). Detecting and analysing spontaneous oral cancer speech in the wild. *arXiv preprint arXiv:2007.14205*.

Hernandez-Mena, C. D. (2019). TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks; shared under the CC BY-NC-ND 4.0 license.

Hernández-Mena, C. D., Meza-Ruiz, I. V., & Herrera-Camacho, J. A. (2017). Automatic speech recognizers for mexican spanish and its open resources [Open Access]. *Journal of Acoustic Modeling*. https://doi.org/10.1016/j.jart.2017.02.001

Hualde, J. I. (2005). *The sounds of spanish*. Cambridge University Press. https://www.cambridge.org/9780521545389

Insight, G. M. (2022). Youtube statistics. https://www.globalmediainsight.com/blog/youtube-users-statistics/

Internet World Stats. (2024). Spanish language usage on the internet [Retrieved July 26, 2024, from https://www.internetworldstats.com/].

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT Press.

Juang, B., & Rabiner, L. (2005, January). Automatic speech recognition - a brief history of the technology development.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14), 7684–7689.

Lipski, J. M. (2008). Central american spanish in the united states. In J. M. Lipski (Ed.), *Varieties of spanish in the united states* (pp. 142–149). Georgetown University Press.

Markl, N. (2022). Language variation and algorithmic bias: Understanding algorithmic bias in british english automatic speech recognition. *Proceedings of the 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*, 521–534. https://doi.org/10.1145/3531146.3533117

Moreno-Fernández, F., & Otero, J. (2007). *Atlas de la lengua española en el mundo*. Real Instituto Elcano-Instituto Cervantes-Fundación Telefónica.

Moro-Velazquez, L., Cho, J., Watanabe, S., Hasegawa-Johnson, M. A., Scharenborg, O., Kim, H., & Dehak, N. (2019). Study of the performance of automatic speech recognition systems in speakers with parkinson's disease. *Interspeech*, 9, 3875–3879.

Nuevo diccionario ejemplificado de chilenismos y de otros usos diferenciales del español de chile. tomos i, ii y iii [Retrieved 2 July 2020, from Universidad de Playa Ancha Sello Editorial Puntángeles]. (2020). https://www.upla.cl/selloeditorial/puntangeles

Parsons, P., Kvale, K., Svendsen, T., & Salvi, G. (2023). A character-based analysis of impacts of dialects on end-to-end norwegian ASR. *The 24rd Nordic Conference on Computational Linguistics*. https://openreview.net/forum?id=boIlH5V81C8

Saon, G., Sercu, T., Rennie, S., & Kuo, H.-K. J. (2016). The ibm 2016 english conversational telephone speech recognition system. https://arxiv.org/abs/1604.08242

Sawalha, M., & Shariah, M. A. (2013). The effects of speakers' gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus. https://api.semanticscholar.org/CorpusID:59726896

Singh, A., & Kadyan, V. (2020). Automatic speech recognition system fortonal languages: State-of-the-art survey. *Archives of Computational Methods in Engineering*, 28. https://doi.org/10.1007/s11831-020-09414-4

Tatman, R. (2017, April). Gender and dialect bias in YouTube's automatic captions. In D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube & H. Wallach (Eds.), *Proceedings of the first ACL workshop on ethics in natural language processing* (pp. 53–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1606

Tellez, E. S., Moctezuma, D., Miranda, S., Graff, M., & Ruiz, G. (2023). Regionalized models for spanish language variations based on twitter [Epub ahead of print]. *Language Resources and Evaluation*, 1–31. https://doi.org/10.1007/s10579-023-09640-9

United Nations. (2024). Spanish official language un [Retrieved July 26, 2024, from https://www.un.org/en/our-work/official-languages].

Wheatley, B., & Picone, J. (1991). Voice across america: Toward robust speaker-independent speech recognition for telecommunications applications. *Digital Signal Processing*, 1(2), 45–63. https://doi.org/https://doi.org/10.1016/1051-2004(91)90095-3

Young, S. (1996). A review of large-vocabulary continuous-speech. *IEEE Signal Processing Magazine*, 13(5), 45–. https://doi.org/10.1109/79.536824

Youtube. (2024). Youtube statistics. https://support.google.com/youtube/answer/7296221hl=en#:~:text=YouTube%2uses%20automatic%20speech%20recognition,Portuguese%2C%20Russian%2C%20and%2Spanish

Zenkovich, A. L. (2018). Particularidades del idioma español en uruguay. https://api.semanticscholar.org/CorpusID:166543415

Zhu, X., Gardiner, S., Rossouw, D., Roldán, T., & Corston-Oliver, S. (2022). Punctuation restoration in spanish customer support transcripts using transfer learning. https://arxiv.org/abs/2205.13961

Zissman, M., Gleason, T., Rekart, D., & Losiewicz, B. (1996). Automatic dialect identification of extemporaneous conversational, latin american spanish speech. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 2, 777–780 vol. 2. https://doi.org/10.1109/ICASSP.1996.543236