Analyzing Multilingual Automatic Speech Recognition Systems Performance

Yetunde E. Adegbegha[1], Aarav Minocha[2], and Renu Balyan[1](✉)

[1] State University of New York, Old Westbury, NY 11568, USA
balyanr@oldwestbury.edu
[2] Great Neck South High School, Great Neck, NY 11020, USA

Abstract. Understanding spoken language, or transcribing the spoken words into text, was one of the earliest goals of computer language processing and falls under the realm of speech processing. Speech processing in itself predates the computer by many decades. Speech being the most important and most common means of communication for most people, is always in need of necessary technology advances. Therefore, in the recent decades there has been great interest in techniques including automatic speech recognition (ASR), text to speech etc. This research is focused around English and the scope needs to be expanded to other languages as well. In this study we explore several open-source ASR systems that offer multilingual (English and Spanish) models. We discuss various models these ASR systems offer, evaluate their performance. Based on our manual observations and using automatic evaluation metrics (the word error rate) we find that Whisper models perform the best for both English and Spanish. In addition, it supports a multilingual model that has the ability to process audio that consists of words from both English and Spanish.

Keywords: Automatic Speech Recognition · Whisper · Vosk · Word Error Rate

1 Introduction

Speech Processing is the study of speech signals and the computer processing methods of these signals in a digital representation [1]. Speech processing is essential in today's technological driven world and creates a more natural human-machine interaction. Speech processing is being used in numerous industries to enhance user experiences and simplify communication. New technologies in the field pave the way for voice activated systems that can enrich our interactions with digital platforms, increase accessibility on the internet for those with disabilities, and help people who speak different languages interact [2]. From language translations to understanding audio through voice biometrics, speech processing is crucial to improving communication in a digital age. The increasing need for new technologies such as smart assistants and real-time translations has made the integration of speech processing in our daily life a necessity in order to foster a new age of modern communication.

Speech recognition technology allows computers to take spoken audio as input, interpret it and generate text (referred to as transcription for the rest of the paper) as an output.

F. Zhao and D. Miao (Eds.): AIGC 2023, CCIS 1946, pp. 191–204, 2024.
https://doi.org/10.1007/978-981-99-7587-7_16

Automatic Speech Recognition (ASR) systems aim at converting a speech signal into a sequence of words either for text-based communication purposes or for device controlling [3]. Research in ASR and speech synthesis has gained a lot of importance and attracted a great deal of attention over the past few decades [4]. Technological curiosity about the mechanisms for mechanical realization of human speech capabilities, and the desire to automate simple tasks inherently requiring human-machine interactions have generated interest in studying the ASR systems [4]. Some of the major growing applications in this field include speech enhancement, speaker recognition and verification, spoken dialog systems, emotion and attitude recognition, speech segmentation and labeling, and audio-visual signal processing. With the number of applications using these voice-based systems, special care needs to be taken while building these systems as failures of ASR systems may result in serious risks to users. For example, in the health domain an ASR system error can pose risk for the patient if the patient is not understood correctly by the ASR [5]. Therefore, further research and a closer investigation is needed to understand the importance of being correctly understood or the consequences of being misunderstood by speech recognition systems [6]. Research has shown that ASR systems exhibit racial bias, and there has been concern over these systems not working equally for everyone [7–9]. Therefore, even though the focus of this study is not identifying the bias in ASR systems towards a particular population, we try to identify if ASR systems perform at the same level for languages other than English, particularly Spanish in this study. There are several well-known ASR systems that have been studied and tested for English, but we found only a few studies that have explored and analyzed the transcripts generated for Spanish using these ASR systems or evaluated the performance of these systems for Spanish [10, 11].

The Goal of the current study was to generate English and Spanish transcriptions from an existing set of recorded videos in the health domain. This study forms a part of a bigger NSF-funded project that is developing a culturally sensitive health intelligent tutoring system (ITS) for the Hispanic population. In order to achieve the said goal, some of the research questions (RQ) that were answered in this study are:

RQ1: What open-source ASR systems exist that can transcribe English as well as Spanish videos?
RQ2: What models within these systems can be used to generate transcriptions for recorded videos based on the performance of the models/systems for the two languages?
RQ3: What evaluation measures can be used to automatically evaluate the system/model's performance?

2 Open-Source ASR Systems

2.1 Whisper

Whisper is a general-purpose, multitasking speech recognition model, trained on 680,000 hours of labeled audio and the corresponding transcripts collected from the internet. This training data constitutes 438,000 hours of English audio and the matching English transcripts; 125,000 hours represents X → English translation data, and the remaining 117,000 hours represent non-English audio and the corresponding transcript, covering 99 other languages. The model was trained using an encoder-decoder transformer [12] as it scales well [13].

2.2 Vosk

Vosk is an open-source and free Python toolkit for offline/online speech recognition. Vosk supports two models - big and small; small models are ideal for limited tasks such as mobile applications. Big models are for high-accuracy transcription and apply advanced AI algorithms. Vosk models provide continuous large vocabulary transcription, zero-latency response with streaming API, reconfigurable vocabulary, and speaker identification. The system can result in poor accuracy due to numerous reasons including bad audio quality, vocabulary mismatch, accent, coding and software bugs [14].

2.3 Kaldi

Kaldi is an open-source toolkit for speech recognition that is written in C + + and is licensed under the Apache License v2.0. More details about Kaldi are available on their website (http://kaldi-asr.org). Kaldi is intended for use by speech recognition researchers and professionals; it is a research speech recognition toolkit that implements many state-of-the-art algorithms. Kaldi has speech activity detection (SAD), speaker identification (SID), language Model (LM), diarization (DIAR) and ASR models with 3 of them being English ASR models [15].

2.4 Julius

Julius is an open-source, high-performance speech recognition decoder for academic research and industrial applications. It supports processing of both audio files and a live audio stream. Julius supports standard language models such as the statistical N-gram model, rule-based grammars, and Hidden Markov Model (HMM) as an acoustic model. Julius in itself is developed as a language-independent decoding program and a recognizer of a language can be developed given an appropriate language and acoustic model for the target language. Julius currently has Japanese and English language/acoustic models [16].

2.5 Mozilla DeepSpeech

DeepSpeech is an open-source Speech-To-Text engine using a model trained by machine learning techniques such as recurrent neural network (RNN) [17]. It uses Google's TensorFlow to make the implementation easier, open and universal [18, 19] A pre-trained English model is available for use [19].

3 Methods

3.1 Data

Eleven short videos of varying length, recorded by a doctor in both English and Spanish were used as the base for the current study. These videos were transcribed to obtain text that was processed further for different purposes using various natural language processing (NLP) techniques. The topic of each video varied but the theme of every video is about cancer survivorship.

3.2 Data Preprocessing

There was no data preprocessing needed for generating the automatic or the human expert transcriptions and the video files were fed into the ASR system or given to the human expert as is for the transcriptions. However, for performing the evaluation of the automatically generated transcription, the data had to be aligned sentence by sentence for both the expert/reference transcriptions and the system-generated/ hypothesis transcriptions as per the requirements of one of the packages (ASR-evaluation) used for the transcription evaluation. This package output was used for an in-depth analysis of the errors produced at the sentence level. However, another package (JiWER) used for evaluating the transcriptions was more flexible and did not require any form of preprocessing on the transcriptions before evaluation or for computing the statistics. More details for the packages are discussed later in the 'Automatic Evaluation' section (see Sect. 5).

3.3 Human Transcription

Human transcriptions were created by a fluent English and Spanish speaker and the transcriptions were later also validated and verified by another speaker fluent in both the languages. The descriptives (number of words and number of sentences) for the transcriptions were generated using SpaCy, an open-source NLP python library are shown in Table 1.

Table 1. Data Descriptives for the English and Spanish Expert/Reference Transcriptions

Transcript	Transcript Description	# of sentences (English/Spanish)	# of words (English/Spanish)
1	Visual Symptoms	5/7	97/79
2	Tamoxifen Side Effects	9/5	217/167
3	Survivorship Care	5/8	243/203
4	SE After Surgery	8/8	198/133
5	PT Side Effects	3/13	162/337
6	PT Breast Cancer Basics	3/9	112/167
7	PT Intro-Mi Guia	2/5	38/68
8	Peripheral Neuropathy	6/8	126/171
9	Osteoporosis	5/6	117/193
10	Depression	9/5	111/131
11	Cardiac symptoms	6/5	120/70

This is to be noted that there are differences between the data descriptives for English and Spanish transcriptions, even though the videos were on the same topics. Some of these differences result due to varying length of videos for the two languages, which leads to different numbers of sentences. In addition, the other differences are caused due

to linguistic differences between the two languages. For example, a word in English may not have a single word equivalent in Spanish but is represented by multiple words or vice versa, which results in a difference between the number of words in the two language transcripts.

3.4 Automatic Transcription

We explored several ASR systems discussed in Sect. 2 for generating automatic transcriptions but only two open-source ASR (Vosk and Whisper) fulfilled the requirements of this study and as a result were used for all the experiments in this study to automatically transcribe our data (videos). Vosk developed by Alpha Cephi supports 27 languages and dialects and Whisper by OpenAI supports 99 languages. We used these ASR systems to transcribe English and Spanish videos. We experimented with the different models that were provided by the two systems. The models finally used for this study were determined based on varying levels of accuracy and speed.

3.5 Experiments

We conducted two experiments in this study, the first experiment was to determine the best open-source ASR for the requirements of our project and the second experiment was to determine the error rates for the ASR-generated transcription to determine the transcription quality. In the first experiment, we implemented several different models provided by the ASR systems to choose the best model. The same dataset was used to test each model's accuracy and speed. In the second experiment, we evaluated and measured the accuracy of the ASR-generated transcriptions using the available error rate evaluation metrics. We manually transcribed the data, (i.e., the videos) to obtain the reference or the expert transcriptions as discussed previously in Subsect. 3.3. We used different ASR evaluation metrics to compute the accuracy of transcripts (i.e., human transcription vs. the automatic transcripts generated by the ASR systems).

Experiment 1: Transcription Models. Several models from Vosk and Whisper were explored to transcribe the data for English and Spanish. For English transcription, Vosk has multiple English models, however, we used 'vosk-model-en-us-0.22' model as this model fulfilled the requirements of our study and was close to what we needed. Whisper supports four English-only models (tiny-en, base-en, small-en, and medium-en). We used the 'model-medium-en' due to its performance and lower error rates.

For Spanish transcription, Vosk supports two models, the 'vosk-model-small-es-0.42' and the 'vosk-model-es-0.42'. We used the 'vosk-model-es-0.42' instead of the 'vosk-model-small-es-0.42', which is a Lightweight wideband model for Android and RPi. It is important to note that the small model is ideal for some limited tasks on mobile applications, while the big models are for the high-accuracy transcription on the server and apply advanced AI algorithms. Since we were not working on mobile applications, we preferred the larger model for this study. Whisper has four models (tiny, base, small, and medium). We decided to use the 'medium' model because it has better punctuations and spellings, and accurately detected the video lengths compared to the other models. While using this model, one has to explicitly state what language the transcriptions are

expected for, because, unlike English, it does not have models trained specifically for Spanish. However, if the language is not explicitly stated, the system detects the language being spoken in the audio or video file and considers the model to be used accordingly.

Experiment 2: ASR Evaluation. The accuracy of each transcription generated by the ASR systems was evaluated through different metrics obtained from the JiWER Python package. These metrics were the Word Error Rate (WER), Match Error Rate (MER), Word Information Loss (WIL), Word Information Preserved (WIP), and Character Error Rate (CER). Another Python package (ASR-evaluation) was also tested for evaluation. This package returned the sentence error rate (SER) and word error rate (WER) but required more data preprocessing and computed information on fewer features. In addition, this package had higher WER as compared to the JiWER package and therefore was not used for initial evaluation. It was also observed that the ASR-evaluation package may be more helpful for deeper analysis for sentence-level transcription evaluation rather than the full transcription.

4 Results

4.1 Transcription Models Performance

For the Spanish transcription, Whisper is the best option to fulfill our project requirements. Vosk and Whisper have similar levels of accuracy with their Spanish models. Whisper also has punctuations to indicate the end of the sentences; whereas, Vosk does not provide punctuations in the transcript and does not have a Spanish model that one can use to include punctuations in the transcripts.

Whisper Spanish (Small vs. Medium model). The Whisper small model is the default model. It has a similar level of accuracy as the Whisper medium model. The manual analysis of the two model outputs indicate that the medium model is a little better for Spanish transcription as compared to the small model. The medium model performed well for all the transcriptions except for Transcript 10 (related to Depression) shown in Fig. 1. The small model for Spanish could perfectly transcribe the name spoken in the video; however, the medium model for Spanish could not transcribe it correctly. It performed poorly than the small model because of differences in paragraphs and space-related issues.

Vosk has multiple English models, yet, only the ‘Vosk-model-en-us-0.22’ was able to generate superior transcriptions as compared to the other models. This model was however not suitable for our needs as it performed well with a generic US-English accent. Whereas, the speaker in our videos has a non-US accent leading the model to perform poorly. As a result, due to the model’s highly inaccurate transcriptions, we decided not to use Vosk for transcribing the English videos. Whisper has four English-only models (tiny-en, base-en, small-en, and medium-en). The default English model is the ‘small-en’ model but we decided to use the model ‘medium-en’ after manually analyzing the transcriptions from the two models. The comparison between the Vosk and Whisper English models and the poor performance of Vosk can be seen clearly in Fig. 2. The difference transcriptions are marked in red, where Vosk indicates incorrect

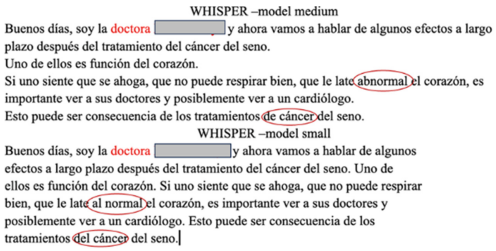

Fig. 1. The transcriptions generated by the Whisper Spanish (medium and small model). The differences in the two models are highlighted and the name of the doctor (which the medium model id not transcribe correctly) has been redacted for confidentiality purposes.

transcription and the Whisper red color transcription represents correct or what was actually spoken in the video. We use the 'Vosk-model-en-us-0.22', a generic US accent model, and the Whisper 'medium-en' model in this example. The Vosk model cannot transcribe accurately because the speaker in the video has a non-US accent. However, the Whisper model can transcribe regardless of the accent. In the first sentence, the speaker introduces herself, which Whisper transcribes correctly as "this is <<Name of the doctor>>.", whereas Vosk transcribes it as "spark oppressed meyer" which is a far cry from what is said. Vosk has no other model that came this close to transcribing our data (see Fig. 2).

In order to reaffirm that it was indeed the non-US accent because of which the Vosk model performed poorly, we transcribed another video (with a generic US accent) randomly selected from the internet. The Vosk model performs well in this case (see Fig. 3) and is able to transcribe words like 'tidbit' and 'inflation' correctly whereas for a non-generic US accent, the model incorrectly transcribed simple words like 'thank you' as 'think'.

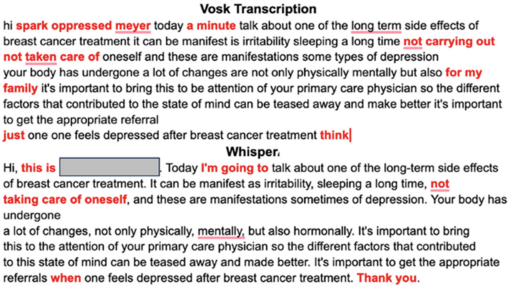

Fig. 2. The transcriptions generated by the Vosk and Whisper English models. The name of the doctor was no where near what it should have been (which has been redacted for confidentiality purposes in Whisper). Red colored words show how poorly the Vosk performs as compared to the Whisper model for English.

Whisper English (Small vs. Medium model). The Whisper English medium and small models both accurately transcribe the data. However, there are minor differences between

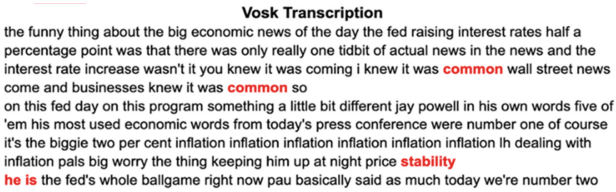

Fig. 3. The transcriptions generated by the Vosk English models for an audio with generic US accent. Vosk does a far better job than it did with no-generic US accent.

the transcriptions returned by the two models as can be seen in Fig. 4. For example, the first line of the medium model ends with 'University'; however, the small model ends with 'at'. This does not impact manual evaluation of the transcription but this results in poor performance during the automatic evaluation. The 'Medium' models overall performs better than the 'Small' model in most transcriptions except it was observed that for Transcript 5 (related to PT Side Effects) it performed differently than the 'small' model because of inconsistent word representations; for example, the small model transcript has the word 'post-menopausal', whereas the medium model transcribes, it as 'postmenopausal'. Both these transcriptions are correct, but the ground truth or a reference transcription will favor the model with a matching word during the automatic evaluation.

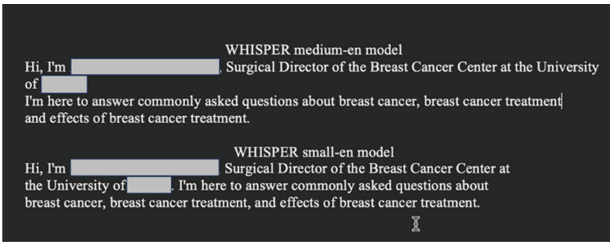

Fig. 4. The transcriptions generated by the Whisper English (medium and small model). There are no differences in the two models transcriptions other than the formatting related, the name of the doctor (the medium model hyphenated the name) and University has been redacted to hold the identity of the person.

Whisper Multilingual Model. Whisper generated all English text with the model 'large' for an audio that contained a mix of both English and Spanish. This was possibly due to the poor audio quality. On the contrary, with the model 'large-v2' in Whisper was able to correctly transcribe an audio file with both English and Spanish (Multilingual). The Whisper multilingual 'large-v2' model performs better than the multilingual 'large' model at transcribing the audio and detecting the languages in the audio. In one of our experiments, the 'large' model transcribed the audio in the language that was dominant rather than the two languages spoken in the audio. It is important to note that when the multilingual audio starts running, the model first detects the language. The 'large-v2'

model has given us consistent, accurate results even though it may also detect the dominant language. The multilingual model also supports language-specific models, therefore depending upon what multilingual model is selected, Whisper transcribes audio in the respective language (English or Spanish in our case) correctly. While using the multilingual model, if the English language is explicitly stated in the command, Whisper transcribes the audio to English regardless of the language in the audio. It first translates the audio from the actual language to English and then transcribes it; it behaves similarly if the Spanish language is explicitly stated in the command.

5 Transcription Evaluation

5.1 ASR Evaluation

The purpose of evaluating ASR systems is to simulate human judgment of the performance of the systems in order to measure their usefulness and assess the remaining difficulties especially when comparing systems; the standard metric of ASR evaluation is the Word Error Rate (WER), which is defined as the proportion of word errors to words processed [3]. The WER is based on how much the output (typically a string of words) called the Hypothesis, returned by the ASR system differs from a reference transcription generated by a human expert. The WER is computed using Eq. (1), where I = number of insertions, D = number of deletions, S = number of substitutions, C = number of correct words and N = number of words in the reference.

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

The Python Jiwer package was used to automatically calculate the WER, Match Error Rate (MER), Word Information Loss (WIL), Word Information Preserved (WIP), and Character Error Rate (CER). The measures are computed with the use of the minimum-edit distance between one or more reference and hypothesis sentences. Although WER is the most popular and commonly used metric to evaluate ASR, it has certain drawbacks [2, 20, 21]. Therefore, many researchers have proposed alternative measures to solve the evident limitations of WER. Andrew et al. [22] introduced Relative Information Lost (RIL) and WIL. WIL value indicates the percentage of words that were incorrectly predicted between a set of ground-truth sentences and a set of hypothesis sentences [23]. WIL is an approximation measure of RIL and is based on HSDI counts. RIL, which is based on Mutual Information (I, or MI), is calculated using the Shannon Entropy H [3]. The CER value indicates the percentage of characters that were incorrectly predicted [23]. The lower the value, the better the performance of the ASR system with a CER of 0 being a perfect score. MER value indicates the percentage of words that were incorrectly predicted and inserted [22, 23]. The lower the value, the better the performance of the ASR system with a MER of 0 being a perfect score. WIP value indicates the percentage of words that were correctly predicted between a set of ground-truth sentences and a set of hypothesis sentences [23, 24]. The higher the value, the better the performance of the ASR system with a WIP of 1 being a perfect score.

5.2 ASR Transcriptions Error Rates

Table 2 shows the evaluation metric results for English Transcription performed by Whisper using the 'medium-en' model. The evaluation metrics in the table includes the metrics returned by the JiWER package namely WER, MER, CER, WIL, and the WIP. The average for each of these metrics for the whole corpus of transcriptions is WER (23.72%), MER (23.49%), CER (4.55%), WIL (35.80%), and WIP (64.20%).

Table 2. The error rates and word information scores for Whisper English model

Transcript	Transcript Description	WER(%)	MER(%)	CER(%)	WIL(%)	WIP(%)
1	Visual Symptoms	21.98	21.50	4.97	34.20	65.80
2	Tamoxifen Side Effects	24.88	24.42	6.14	37.80	62.20
3	Survivorship Care	19.17	19.01	3.97	28.86	71.14
4	SE After Surgery	19.89	19.89	4.12	30.60	69.39
5	PT Side Effects	25.48	25.16	4.87	37.36	62.64
6	PT Breast Cancer Basics	20.75	20.75	3.76	32.76	67.24
7	PT Intro- Mi Guia	40.54	40.54	5.81	59.12	40.88
8	Peripheral Neuropathy	22.58	22.22	3.99	33.23	66.77
9	Osteoporosis	19.30	19.30	3.03	30.61	69.39
10	Depression	23.56	23.15	4.74	35.01	64.99
11	Cardiac symptoms	22.81	22.41	4.66	34.21	65.79

Table 3 shows the same evaluation metrics as in Table 2 but for Spanish transcription for the Whisper using the 'medium' model and Vosk using the 'Vosk-model-es-0.42'. We observed that Whisper outperforms the Vosk model in all the transcriptions accuracy except for Transcripts 7 and 10. Vosk was better due to formatting (paragraphs and spaces) after we compared it with reference transcript. These numbers will change if the formatting in the reference transcript changes. However, in Transcript 10, Vosk in addition to the formatting issues, transcribed the name of the doctor correctly, but Whisper could not.

5.3 WER Related Challenges

The fundamental problem with the WER is that it weighs every word equally. For example, a determiner and an adjective will be treated the same, even though as humans we know that not every word is important and some errors matter more than others. Because the context determines some of these factors, it is difficult to develop a test that can be broadly applied. In addition to ignoring the importance of words, the WER does not give any partial credit. Even if a mis-transcribed word mismatches by just one character, WER treats it as incorrect or a mismatch. The WER does not account for speaker labels and punctuations, which may be important in some cases. Another issue to

Table 3. The error rates and word information scores for Spanish transcriptions (Whisper vs Vosk models)

Transcript	Transcript Description	WER (%) Whisper/ Vosk	MER (%) Whisper/ Vosk	CER (%) Whisper/ Vosk	WIL (%) Whisper/ Vosk	WIP (%) Whisper/ Vosk
1	Visual Symptoms	22.08/31.17	21.52/29.63	5.99/8.87	35.17/46.59	64.83/53.41
2	Tamoxifen Side Effects	29.70/41.21	28.49/38.86	11.30/17.68	45.09/58.20	54.90/41.80
3	Survivorship Care	21.89/27.36	21.57/26.44	5.22/7.65	33.32/41.18	66.68/58.82
4	SE After Surgery	25/28.79	24.26/28.15	10.27/11.55	36.72/46	63.28/54
5	PT Side Effects	21.92/26.73	20.98/25	7.34/9.86	32.81/37.40	67.19/62.60
6	PT Breast Cancer Basics	22.29/23.49	21.26/22.54	7.46/9.98	32.30/36.37	67.70/63.63
7	PT Intro- Mi Guia	**27.27/25.76**	**25.71/24.29**	**5.99/7.91**	**40.62/39.20**	**59.38/60.80**
8	Peripheral Neuropathy	26.04/32.54	25/30.22	9.25/12.68	38.26/45.46	61.74/54.54
9	Osteoporosis	25.93/33.33	25.52/32.14	17.04/18.40	35.21/45.90	64.79/54.10
10	Depression	**33.85/26.92**	**33.08/26.12**	**9.97/7.80**	**51.26/42.01**	**48.74/57.99**
11	Cardiac Symptoms	27.94/29.41	26.76/27.40	8.38/9.14	42.37/43.41	57.63/56.59

consider for accuracy is that a verbatim transcript is likely to include many meaningless words such as “umms”, “uhs”, duplicates and false starts, which do not add anything meaningful to the text [25]. Some of the high error rates seen in Tables 2 and 3 can be attributed to several of WER-related issues as were also observed in our transcripts during the manual analysis of the transcripts.

6 Discussion

In this study, we conducted 2 experiments to answer the three research questions (RQ1–RQ3) discussed in the Introduction (Sect. 1).

6.1 RQ1: What Open-Source ASR Systems Exist that can Transcribe English as well as Spanish Videos?

Several open-source ASR systems including DeepSpeech, Julius, Kaldi, Vosk and Whisper were explored in the study and after some initial research and analysis we selected

two open-source ASR systems namely Whisper and Vosk for further experiments in this study. Both these systems can transcribe both English and Spanish audio. We did not continue experiments with DeepSpeech, as it is no longer supported and there have been no new versions or releases since 2020 [19]. Julius has support for English, Thai, Chinese, Korean but no popular version of Spanish has been found. A paper cited only once since 2014 used it to propose the Spanish version. Lack of available Spanish data for our population was another reason for not being able to train our own model for Julius and Kaldi. Whisper outperforms Vosk in both the languages English as well as Spanish. In addition, the Whisper transcriptions are more readable than Vosk because Whisper models return punctuations and capitalization as humans.

6.2 RQ2: What Models within these Systems can be Used to Generate Transcriptions for Recorded Videos based on the Performance of the Models/systems for the Two Languages?

All the English and Spanish models for Whisper and Vosk can generate transcription for recorded videos. However, each model's accuracy level varies; the Vosk English models could only partially transcribe the recorded videos as they are trained with US accents. The requirements for our project need an ASR that considers all accents, as the target population for the project is Hispanic individuals who are less likely to have a generic US accent. Whisper, however, can accurately transcribe videos regardless of the speaker's accent. The Vosk and Whisper Spanish models were able to transcribe the videos accurately. We decided that the Whisper 'medium-en' model for English and the Whisper 'medium' model for Spanish were the best available options for transcribing the videos.

6.3 RQ3: What Evaluation Measures can be Used to Automatically Evaluate the System/model's Performance?

The standard metric for the ASR evaluation is the WER. However, other metrics can be used to evaluate the systems, as WER also has flaws and is not perfect. There are other metrics like the WIL, MER, CER, and WIP. We used the Python Jiwer package to automatically calculate these metrics to compare and determine the best models/systems.

6.4 Limitations

Like any study, this study also has a few limitations that need further research and some future work. 1) So far, Whisper provides the most suitable transcription that best serves our purpose. Although Whisper supports 99 languages, based on our analysis the English models are better trained than the Spanish. In this study we focused mainly on the models that independently support either English or Spanish. However, our future work seeks to find a multilingual ASR because the project's target population (Hispanic) is bilingual (who speak both Spanish and English). Whisper has a large model, which is primarily multilingual, so we will be exploring this model in our future work. 2) Even though the WER is the standard ASR evaluation metric, it has considerable issues, and there have

been criticisms against solely relying on it [2, 21]. The accuracy of the ground truth (i.e., manual transcript) also greatly affects the WER. Humans make errors, the manual transcript could be incorrect, but the WER works with the assumption that the ground truth is perfect. Even while working under the assumption that the manual is perfect, the grammar affects the error rate. For example, "Dr." and "Doctor" is correct; however, if the ground truth uses "Doctor" and the hypothesis uses "Dr." the WER calculates that as an error, but they are both right. We will need to explore better and efficient measures to evaluate ASR accuracy in our future studies.

7 Conclusions

A lot of ASR systems exist where the research focus has been mainly English, and minimal research is available for multilingual ASR systems. Therefore, this paper explored the ASR systems with the focus on their capability of handling multiple languages as well as multilingual audio. The ASR systems explored in the study either did not support Spanish (language of our interest for the study) or did not perform well. The ASR system by Whisper was the only system that supported both the languages (English and Spanish) and performed well. In addition, Whisper supports a model that is also capable of handling mixed audio, which is our anticipated data from the target population. The evaluation metrics mostly are based of WER and we discussed several challenges encountered while using WER for evaluation, indicating there needs to be more research for ASR evaluation, better and more efficient techniques are needed in this area.

Acknowledgment. This work was supported by grants from the National Science Foundation (NSF; award# 2131052 and award# 2219587). The opinions and findings expressed in this work do not necessarily reflect the views of the funding institution. Funding agency had no involvement in the conduct of any aspect of the research.

References

1. Garza-Ulloa, J.: Introduction to cognitive science, cognitive computing, and human cognitive relation to help in the solution of artificial intelligence biomedical engineering problems. In: Applied Biomedical Engineering Using Artificial Intelligence and Cognitive Models, pp. 39–111 (2022)
2. Kong, X., Choi, J.Y., Shattuck-Hufnagel, S.: Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 5810–5814. IEEE (2017)
3. Errattahi, R., El Hannani, A., Ouahmane, H.: Automatic speech recognition errors detection and correction: a review. Procedia Comput. Sci. **128**, 32–37 (2018)
4. Juang, B.H., Rabiner, L.R.: Automatic Speech Recognition–a Brief History of the Technology Development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara (2005). https://doi.org/10.1016/B0-08-044854-2/00906-8
5. Topaz, M., Schaffer, A., Lai, K.H., Korach, Z.T., Einbinder, J., Zhou, L.: Medical malpractice trends: errors in automated speech recognition. J. Med. Syst. **42**(8), 153–154 (2018)

6. Mengesha, Z., Heldreth, C., Lahav, M., Sublewski, J., Tuennerman, E.: "I don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on African Americans. Front. Artif. Intell. **4**, 169. (2021)
7. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., et al.: Racial disparities in automated speech recognition. Natl. Acad. Sci. **117**(14), 7684–7689 (2020)
8. Harwell, D.: "The Accent Gap". The Washington Post. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/ (2018). Last accessed 14 Aug 2023
9. Tatman, R.: Gender and dialect bias in YouTube's automatic captions. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pp. 53–59 (2017)
10. Zea, J.A., Aguiar, J.: "Spanish Políglota": an automatic Speech Recognition system based on HMM. In: 2021 Second International Conference on Information Systems and Software Technologies (ICI2ST), pp. 18–24. IEEE (2021)
11. Hernández-Mena, C.D., Meza-Ruiz, I.V., Herrera-Camacho, J.A.: Automatic speech recognizers for Mexican Spanish and its open resources. J. Appl. Res. Technol. **15**(3), 259–270 (2017)
12. Vaswani, A., et al.: Attention is all you need. Advances in neural information processing systems (2017)
13. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision. ArXiv (2022)
14. Vosk Documentation. https://alphacephei.com/vosk/. Last accessed 14 Aug 2023
15. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society (2011)
16. Lee, A., Kawahara, T.: Recent development of open-source speech recognition engine Julius. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.131–137. Asia-Pacific Signal and Information Processing Association (2009)
17. Hannun, A., et al.: Deep Speech: Scaling up end-to-end speech recognition (2014)
18. DeepSpeech Documentation. https://deepspeech.readthedocs.io. Last accessed 14 Aug 2023
19. DeepSpeech Python Library. https://pypi.org/project/deepspeech. Last accessed 14 Aug 2023
20. Maier, V.: Evaluating ril as basis for evaluating automated speech recognition devices and the consequences of using probabilistic string edit distance as input. 3rd year project. Sheffield University (2002)
21. Szyma´nski, P., et al.: WER we are and WER we think we are. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3290–3295. Online. Association for Computational Linguistics (2020)
22. Morris, A.C.: Maier, V., Green, P.: From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. Interspeech (2004)
23. TouchMetrics Homepage. https://torchmetrics.readthedocs.io. Last accessed 14 Aug 2023
24. Morris, A.C.: An information theoretic measure of sequence recognition performance. IDIAP (2003)
25. Kincaid,J: Challenges in Measuring Automatic Transcription Accuracy. https://medium.com/descript/challenges-in-measuring-automatic-transcription-accuracy-f322bf5994f (2018)