Lexical Analysis of Automatic Transcriptions Using Speech-to-Text Services: A Statistically Evaluated Case Study

Venilton FalvoJr
University of São Paulo (ICMC-USP)
and DIO
falvojr@usp.br

Anderson da Silva Marcolino
Federal University of Paraná (UFPR)
anderson.marcolino@ufpr.br

Diego Renan Bruno
University of São Paulo (ICMC-USP)
diego_renan_bruno@hotmail.com

Catherine Helen Martins Falvo
University of Araraquara (UNIARA)
chmfalvo@uniara.edu.br

Fernando Santos Osório
University of São Paulo (ICMC-USP)
fosorio@icmc.usp.br

Ellen Francine Barbosa
University of São Paulo (ICMC-USP)
francine@icmc.usp.br

Abstract

This paper introduces Speech2Learning, an innovative architecture designed to leverage Speech-To-Text (STT) technology to enhance the accessibility of Learning Objects (LOs). Stemming from a recognized gap in prior Systematic Mapping, the primary objective of this architecture is to simplify the development of flexible educational solutions. In a collaborative endeavor with Brazilian EdTech DIO, we instantiated Speech2Learning as a Proof of Concept (PoC) to subtitle video lessons on their e-learning platform. This PoC was essential to obtain valuable insights for a more comprehensive Case Study. Therefore, we performed a lexical similarity analysis on the automatic transcriptions generated by leading STT providers in Portuguese, English and Spanish. Finally, we carried out a rigorous Statistical Analysis to evaluate the quantitative data from the Case Study. Our findings highlight the potential of Speech2Learning to promote the accessibility of LOs, as well as the relevance of continued research to increase the accuracy of STT services.

Keywords: Speech-to-Text (STT), Learning Objects (LOs), Case Study, Lexical Analysis, Statistical Evaluation.

1. Introduction

In the digital age, education is continually evolving with emerging technologies that have the potential to reshape traditional pedagogical approaches. Against this background, the Automatic Speech Recognition (ASR), represented in this paper by Speech-To-Text (STT), stands as a powerful tool. The STT concept can not only enhance the accessibility of content with transcripts and subtitles, but also promotes a move towards more inclusive education. Aligning with Homburg et al. (2019), we highlight the significant potential of STT in assistive technologies in the deaf community, a perspective reinforced in our previous Systematic Mapping (SM) featuring text-based sign language avatars (FalvoJr et al., 2020).

According to Fleischmann et al. (2021), the rise of remote learning, accelerated by events such as the COVID-19 pandemic, has driven the search for innovative methods for creating and sharing educational content. In this context, the STT has consolidated itself as a promising tool, particularly in collaborative environments or online conferences. STT-based solutions play a vital role in breaking language barriers, optimizing communication between speakers of different languages. This argument is reinforced by Homburg et al. (2019), which presents the relevance of voice translation into sign languages in order to promote the inclusion of the deaf community in the teaching-learning process.

Despite its potential, the STT faces several obstacles and research challenges. Koenecke et al. (2020) highlights some of them, pointing to racial disparities and the subtleties of linguistic characteristics, such as accents and regional peculiarities. These findings reinforce the relevance of promoting STT-based solutions that are truly inclusive and that address a broader spectrum of sociolinguistic considerations in their design and implementation.

Mayer and Fiorella (2021) state that the use of disruptive technologies, such as STT, is essential to expand the reach of Learning Objects (LOs) to a greater diversity of learners, promoting more accessible educational content. In practice, Parakh et al. (2022) describes LOs as reusable digital units that are often integrated into open-source initiatives, playing a crucial role in shaping adaptable, democratic and contextualized teaching-learning experiences.

These recent perspectives corroborate the insights of our SM, emphasizing the importance of technological innovations in the sign language teaching-learning process, in addition to highlighting relevant gaps. Briefly, we analyzed 185 primary studies, where we noticed the lack of design patterns and best practices in the solutions, which compromises the efficient sharing and reuse of LOs (FalvoJr et al., 2020). Faced with these gaps and the aforementioned emerging trends, we designed the *Speech2Learning* Architecture. An abstraction that propose development guidelines for creating SST-based solutions, promoting greater accessibility of their LOs, especially the audible ones.

To demonstrate the potential of our architecture, we implemented a Proof of Concept (PoC) in partnership with

URI: https://hdl.handle.net/10125/107023
978-0-9981331-7-1

DIO, a Brazilian EdTech that not only boasts over one million registered students on its e-learning platform (*https://dio.me*) but also connects these learners with big companies. This PoC instantiated the *Speech2Learning* to provide subtitles, derived from transcripts, for a selected set of video lessons in the three languages: Portuguese, English, and Spanish. However, we found that the automatic transcripts generated by the DIO's cloud provider fell short of our anticipated quality standards.

For this reason, experts in each language reviewed the automatic transcripts using our PoC, ensuring the reliability of these transcripts. Although onerous/expensive, this process resulted in high-quality reference transcripts, providing valuable artifacts for a subsequent Case Study. In turn, the Case Study aimed to obtain quantitative data through lexical analysis, measuring the similarity between the reference transcripts (reviewed in the PoC) and the automatic transcripts from several STT providers. The quantitative data was analyzed to determine the existence of statistically significant differences between the providers and languages evaluated.

In particular, our research presents three main contributions around the *Speech2Learning* Architecture: (i) an industry-applied PoC that generated insights and produced expert-reviewed high-quality transcripts for video lessons in Portuguese, English, and Spanish; (ii) a Case Study evaluating STT services from leading providers—Amazon, Google, IBM, Microsoft (Gartner, 2023), and OpenAI—using lexical analysis to quantitatively assess automatic transcripts; and (iii) a Statistical Analysis examining hypotheses to determine the statistical significance of transcription quality among the evaluated providers for each language.

This paper is structured as follows: Section 2 provides an overview of *Speech2Learning* and its PoC, which generated the reference transcripts explored in the Case Study. The Section 3 details the Case Study, which compares reference and automatic transcripts using lexical analysis methods. Based on the quantitative data collected, we present our statistical analysis and discussions in Section 4. Finally, Section 5 summarizes our findings and their implications.

2. Proof of Concept

A PoC is a practical demonstration that validates if a system, method or idea is feasible in the real world. It often serves as a prototype or preliminary model that assists in evaluating the potential of the proposed solution (Sommerville, 2015). In this study, our PoC is a *Speech2Learning* Instance implemented as a REST API, which was designed for transcribe video lessons, enabling subtitling in Portuguese, English and Spanish.

To support both the PoC and the subsequent Case Study, we collaborated with DIO, an emerging Brazilian EdTech. They provided us with video lessons from their curriculum, creating a valuable link between the industry and our research. We also gained access to a portion of DIO's cloud infrastructure, including Google's STT service. Furthermore, the company's language experts contributed by reviewing the automatic transcripts generated by our PoC, the *Speech2Learning* Instance.

2.1. The *Speech2Learning* Architecture

Speech2Learning is designed to encourage the development of well-structured solutions that aim to include the greatest number of students in the teaching-learning process through more accessible content. In practice, this architecture proposes enriching LOs with accessibility-centric data, such as transcripts, closed captions, subtitles, translations, licensing, version control, and more. Indeed, a simple transcript can enable very interesting features, such as integration with text-based sign language avatars.

Technically, the *Speech2Learning* is an adaptation of the Clean Architecture, which is a design paradigm that favors building systems that are easily maintainable, testable, and extendable, while remaining agnostic to particular frameworks, user interfaces, and databases (Martin, 2017). According to Martin (2021), Clean Architecture itself is a comprehensive approach that integrates some of the major references in Software Engineering over the last decades. Based on this philosophy, the layers of *Speech2Learning* are designed to be modular, testable, and centered on providing accessible LOs (Figure 1):

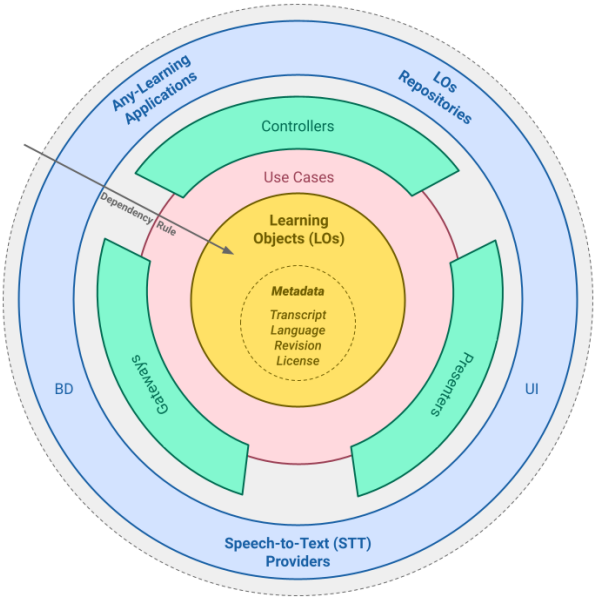


Figure 1: *Speech2Learning* Architecture

- **LOs (Yellow):** This layer is the core of our architecture, where the LOs are Entities that encapsulate critical business rules of the educational domain aiming at accessibility. The LOs can be a class with methods, or a set of data structures and functions. The paradigm type doesn't matter, as long as the LOs are generic and can be used by several applications. *Speech2Learning* proposes that all LOs have metadata, a structure reserved for any kind of information that promotes LOs accessibility: transcript, language, license, revision etc.
- **Use Case (Red):** The code in the Use Case layer contains application-specific business rules. This layer

encapsulates and implements all system features, the highest level operations that users perform. A Use Case orchestrate the data flow among entities, directing them to use their critical business rules to achieve their goals.

- **Adapters (Green):** Responsible for converting data into a convenient format, considering the needs of the layers it interfaces with. Gateways, for example, will adapt data obtained by the STT Providers into a suitable format for the Use Case and LOs layers.
- **Infrastructure (Blue):** The outermost layer of *Speech2Learning* is composed of frameworks, libraries, and other solutions. It is in this layer where all the implementation details are evident. As discussed by Martin (2017), the Database is just a detail, as is the Web. So it doesn't matter what type of learning application (d-learning, e-learning, m-learning etc) or what STT Provider is used, as the inner layers are independent of these "details".
- **Main and Configuration (Gray):** Although not a formal layer, it establishes all connections between interfaces and their concrete implementations. It's also responsible for the system's execution.

Additionally, it's crucial to clarify the "Dependency Rule": source code dependencies should only point inwards towards high-level policies, never outwards. In summary, the *Speech2Learning* favors adaptability, technological independence, testability and a structural guidance for the creation of more inclusive educational solutions.

To conclude, it's important to mention that although *Speech2Learning* was inspired by the results of a SM study focused on sign languages (FalvoJr et al., 2020), our architecture is generic and seeks to promote accessibility to any kind of LOs. However, in this paper's context, audible LOs will be the focus. In this regard, we present a *Speech2Learning* Instance below, applied as a PoC, for the transcription of video lessons in an online educational platform.

2.2. Implementation of the PoC: A *Speech2Learning* Instance

For the development of the PoC, we created an instance of *Speech2Learning* as a REST API to transcribe video lessons from DIO's e-learning platform. Utilizing the metadata concept inherent in the *Speech2Learning* Architecture, we converted the transcripts into subtitles for a series of selected courses, thereby enhancing the accessibility of these LOs.

As shown in Figure 2, the REST API adheres to the *Speech2Learning* Architecture guidelines. The components' color scheme matches that of the layers in Figure 1, illustrating a compliant implementation within the predefined logical/structural boundaries for the REST API.

Note that the PoC representation makes it clear that there are two Use Cases related to transcribing videos: *Create* and *Review*. This view is interesting because it clarifies the features implemented by the PoC. Furthermore, the Entity was modeled as an *Audible LOs*, since we are dealing with video resources. To conclude, details about the implementation

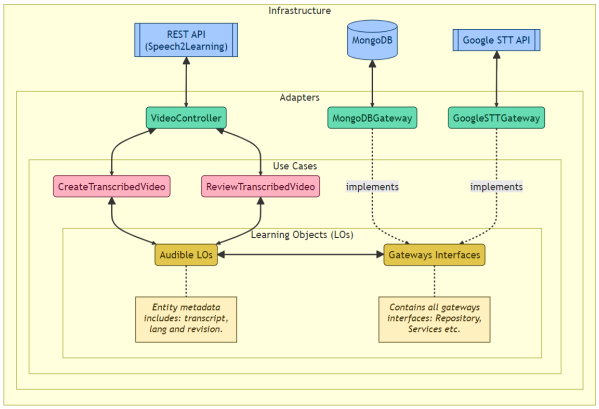


Figure 2: *Speech2Learning* Instance (PoC): REST API

are also clearly exposed in the Infrastructure; for example, our entrypoint is a *REST API*, *MongoDB* is the database and *Google STT API* (external API) is the STT provider.

To offer a more technical perspective, the PoC initializes on-demand through an HTTP POST request to a reactive REST API. This endpoint triggers an asynchronous workflow to extract the video's audio, thereby optimizing bandwidth consumption to invoke *Google STT API*. Upon response to this automatic transcription request, the transcript is stored as metadata in *Audible LOs*, making it eligible for review.

To revise the transcript of an existing LO, an HTTP PUT request can be made which will update this resource by versioning your transcripts using the metadata. Note that *Speech2Learning* only defined the guidelines for creating STT-based solutions to promote accessibility of LOs, without imposing design decisions, technologies or security aspects.

In practice, access to the PoC was restricted to DIO's Education Team, ensuring that their LOs (video lessons) were created and reviewed securely. We emphasize that the automatic transcriptions required extensive review by company's language experts, highlighting the need to optimize the transcription process and explore other STT services.

2.3. Dataset Rationale for the Case Study

Motivated by the PoC's findings, we decided to expand our insights through a Case Study. This research initiative is designed to further analyze the nuances between different STT providers. We chose highlights (10-30 seconds) from 15 video lessons, which were reviewed in partnership with DIO's language experts. The quantity and duration of these videos were carefully planned to manage STT service costs and enable a future qualitative Survey on automatic transcriptions.

For a more robust analysis, we sought to diversify the languages and teachers featured in the videos, in order to capture different accents and regional dialects. Out of the 15 selected videos, we have: 5 in Portuguese (Brazil), 5 in English, and 5 in Spanish. Table 1 details the linguistic and technical characteristics of the extracted audios (highlights). This diverse dataset will serve as a benchmark, allowing us to critically analyze the performance of other STT providers.

ID	Lang	Accent	Gender	Lesson Topic	Time
1	pt-BR	BRA	M	Android Apps	0:17
2	pt-BR	BRA	M	SCRUM	0:26
3	pt-BR	BRA	F	Selenium WebDriver	0:23
4	pt-BR	BRA	F	Blockchain	0:20
5	pt-BR	BRA	M	Hybrid Kernel	0:24
6	en-US	BRA	F	Transit Visa	0:29
7	en-US	USA	Both	Job Interview	0:16
8	en-US	BRA	F	Job Opportunities	0:20
9	en-US	BRA	F	Servant Leadership	0:15
10	en-US	BRA	M	Goroutines	0:15
11	es-AR	ARG	F	Programming Logic	0:12
12	es-AR	ARG	F	Programming Languages	0:21
13	es-AR	ARG	F	Python Data Types	0:14
14	es-AR	ARG	F	Python Hello World	0:17
15	es-AR	ARG	F	Python String Slicing	0:26

Table 1: Summary of Case Study Dataset (Audios)

Notice that this dataset has undergone a rigorous audio quality control process, a standard protocol for all educational content offered by DIO. To validate these claims and ensure transparency, we've made available a public folder[1] that contains all audio samples (Table 1) and their respective transcripts. This dataset meet or exceed the following industry-standard minimum criteria: dual audio channels, a 44.1 kHz sample rate and 16 bit precision. Such technical aspects not only guarantee excellent sound quality, but also contribute to more accurate automatic transcriptions.

In this sense, conducting a Case Study is an excellent option to expand our PoC and guide deeper analyses and discussions. Our PoC identified the need to reduce the rework required in automatic transcripts and thus improve the efficiency of the transcription process. This progression from PoC to Case Study reflects a common approach in research that allows for iterative development (Runeson & Höst, 2009). The dataset compiled in this PoC makes it possible to analyze the quality of automatic transcripts across various STT providers, using the reference transcripts reviewed by DIO's language experts as a reliable benchmark. This complete analysis will be presented below.

3. Case Study

This section presents the planning, methods, and results of a Case Study conducted to evaluate and identify the most accurate STT provider for automatic transcription. This is an empirical investigation, relying on inductive reasoning and field research. Unlike experimental studies, case studies gather information from various sources through diverse data collection techniques (Sommerville, 2015).

First, we selected the main STT providers on the market (Gartner, 2023): Amazon, Google, IBM, Microsoft and OpenAI. Then, DIO provided access to the files of the 15 video classes rationalized in the dataset, which has content in Portuguese, English and Spanish. In this sense, we made cuts in the videos to delimit the scope of this Case Study. Using the *Speech2Learning* Instance (PoC), we generated reference transcripts that were reviewed by language specialists from the company. These will serve as a benchmark when compared with the automatic transcripts generated by each provider[2].

[1]Case Study Dataset (Audios): https://bit.ly/S2L-Audios

[2]STT Providers Services (Colab): https://bit.ly/S2L-STTServices

Measuring the similarity between texts is a common practice in academic research, usually supported by Majumdar (2022) lexical analysis methods. Therefore, in order to compare our automatic and reference transcripts, we need to define how we will extract our quantitative data using lexical analysis techniques. In our context, some of the most interesting alternatives are: Cosine Similarity – CS (Lahitani et al., 2016; Mohana & Suriakala, 2018; Ristanti et al., 2019), Jaccard Index – JI (Manalu et al., 2019; Sulaiman & Mohamad, 2012) and Levenshtein Distance – LD (Sugiarto et al., 2020; Zhang et al., 2017).

For our Case Study, these three methods were applied and have great relevance in our results. In practice, we selected 15 short video lessons for transcription, 5 in each language: Portuguese, English and Spanish. First, reference transcripts were generated and reviewed using the *Speech2Learning* Instance. Subsequently, each video lesson will be automatically transcribed by the STT services of each target provider of this Case Study. With this, it will be possible to apply the lexical analysis methods of CS, JI and LD.

CS is a metric used to measure the similarity of two vectors. Specifically, it measures the similarity in direction or orientation of vectors, ignoring differences in their magnitude or scale. Both vectors must be part of the same inner product space, which means they must produce a scalar by multiplying the inner product. The similarity of two vectors is measured by the cosine of the angle between them (Lahitani et al., 2016; Mohana & Suriakala, 2018; Ristanti et al., 2019). The CS metric measures the cosine of the angle between two n-dimensional vectors projected onto a multidimensional space, it ranges from "0" to "1". A value closer to "0" indicates less similarity, while a score closer to "1" indicates more similarity.

JI is defined as an intersection of two texts divided by the union of these documents, measuring the similarity between two sets of data. That is, it can be expressed as the number of common words over the total number of words in the two texts or documents. The JI of two documents ranges from 0 to 1, where 0 means no similarity and 1 means complete overlap (Majumdar, 2022). JI is calculated by dividing the number of observations in both sets by the number of observations in each set. That is, the JI can be calculated as the intersection size divided by the union size of two sets.

Finally, LD is a string metric for measuring the difference between two strings. Basically, the Levenshtein distance between two words is the minimum number of single-character edits (ie insertions, deletions, or replacements) required to change one word into another. It is named after Vladimir Levenshtein, who considered this distance in 1965 (Sugiarto et al., 2020; Zhang et al., 2017). LD comparison is usually performed between two words. It determines the minimum number of single-character edits required to change one word to another. The greater the number of editions, the more the texts differ from each other. An edit is defined by inserting a character, deleting a character, or replacing a character.

For the execution, a total of 15 transcriptions of short videos were submitted to each of the three selected methods

(CS, JI and LD)[3]. There were five transcriptions for each language: Portuguese, English, and Spanish. The results obtained through the methods of lexical similarity analysis applied to our data set formed by texts generated with the STT tools are presented next.

3.1. Case Study Results

When analyzing the graphs in Figures (4a , 4b, 4c, 5a, 5b, 5c, 3a, 3b and 3c) in detail, we can see that the Cosine, Jacard and Levenshtein metrics provide, in essence, similar feedbacks in relation to each tool compared in our study. In all evaluated metrics, OpenAI's Whisper tool showed marginally better performance in its tests. For example, on Levenshtein's distance metric, where a smaller value represents a better result, OpenAI's Whisper stands out with its results. In addition, it is evident that, in the evaluation based on the cosine and the Jacard index, where values closer to 1 indicate a better performance, Whisper from OpenAI also stands out.

Essentially, we can conclude that, across different metrics, we get a consistent pattern in the results, indicating that OpenAI's Whisper tool has a significant advantage over Amazon, Google, and IBM tools. It is important to emphasize that, in the next section of this study, we will formalize this hypothesis, which was presented based on these conclusions, in a statistical way, demonstrating the relationship between the best tools for converting speech into text.

3.1.1. Cosine Similarity (CS): Measures the similarity between two vectors. In this case, we can consider each generated transcript as a vector of words (Bosker, 2021; Kumar & Renuka, 2023; Sasu, 2019; Tanberk et al., 2021).

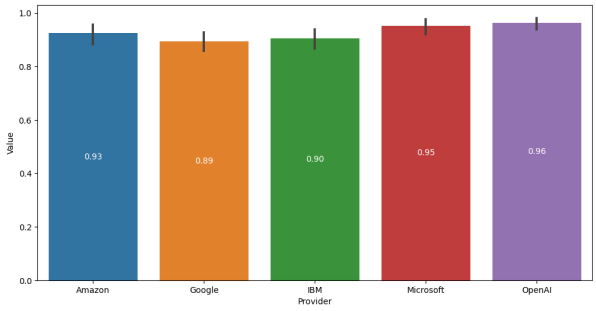
3.1.2. Jaccard Index (JI): Is a metric that measures the overlap between two sets. In this case, we can consider each generated transcript as a set of words (Bosker, 2021; Hapke et al., 2019; Kumar & Renuka, 2023; Noel, 2020; Sasu, 2019).

3.1.3. Levenshtein Distance (LD): Levenshtein's measure is based on the minimum number of operations required to transform one string into another, with the allowed operations being insertion, deletion and replacement of a character. The smaller the value of the Levenshtein measure, the more similar are the two sequences (Bosker, 2021; Hapke et al., 2019; Hasan et al., 2020; Kumar & Renuka, 2023; Noel, 2020; Sasu, 2019).

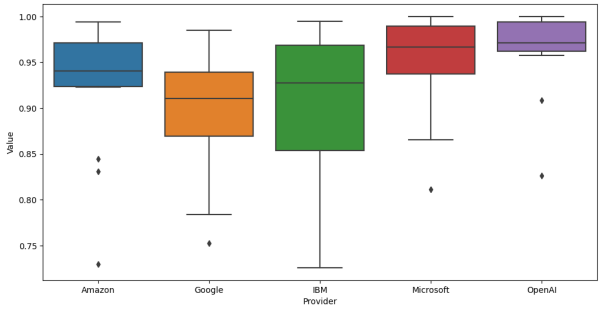
Notice that the cosine similarity and the Jaccard similarity are measures of similarity that range from 0 to 1, where 0 indicates no similarity and 1 indicates that the two vectors are exactly the same. In theory, you should never see a value above 1 for any of these metrics.

We can see in the graph of Figure 4c and graph of Figure 3c that the values are above 1. What could be happening here is a feature of the kernel density graph (KDE). KDE is a way to estimate the probability density function of a random variable. Notice that while the integral (area under the curve) of a probability density function always equals 1,

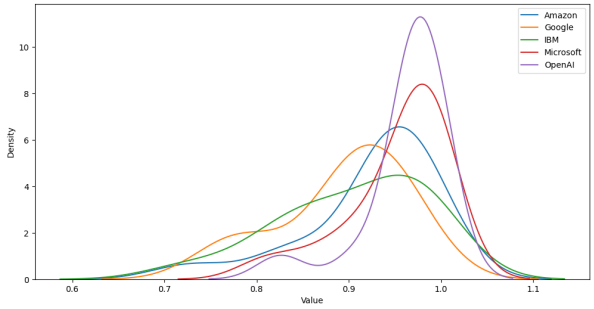
[3]Lexical Analysis (Colab): https://bit.ly/S2L-CaseStudy



(a) Cosine Similarity Bar Plot.



(b) Cosine Similarity Box Plot.



(c) Cosine Similarity KDE Plot.

Figure 3: Plots of Cosine Similarity.

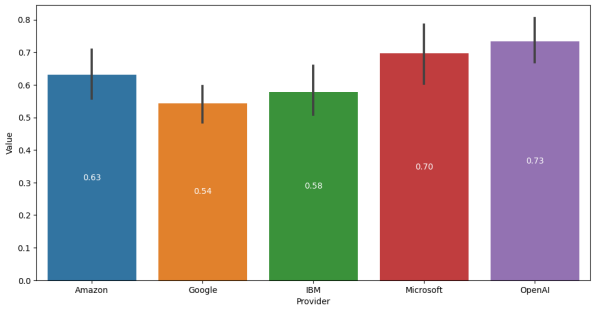
individual values in the probability density function can be greater than 1.

Although the similarity metrics are limited to the [0, 1] range, the values in the kernel density graph can be greater than 1. This does not mean that the metrics are crossing their limits; rather it is a property of KDE and the way it estimates the probability density function.

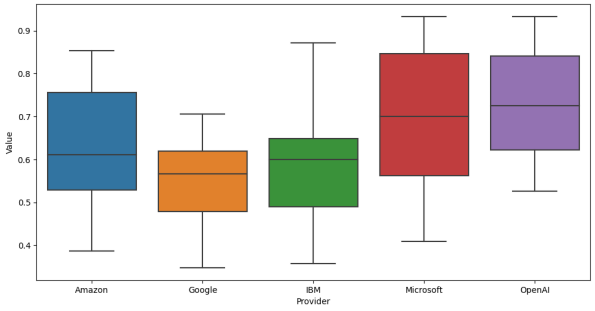
3.2. Language Analysis

Through the graphs of Figures 6a, 6b and 6c we can observe the behavior of the SST tools for the languages: English, Spanish and Portuguese. In the results using CS and JI, where values closer to 1 represent the best responses of the STT system, note that the OpenAI tool had better results compared to Amazon, Google, IBM and Microsoft tools. The best results presented by the OpenAI tool were better in the three applied languages.

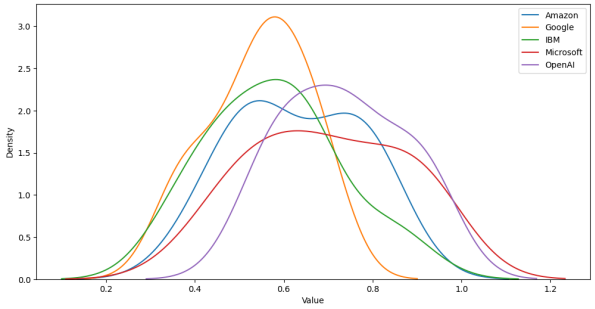
Highlighting that the diversity of languages in STT



(a) Jaccard Index Bar Plot.

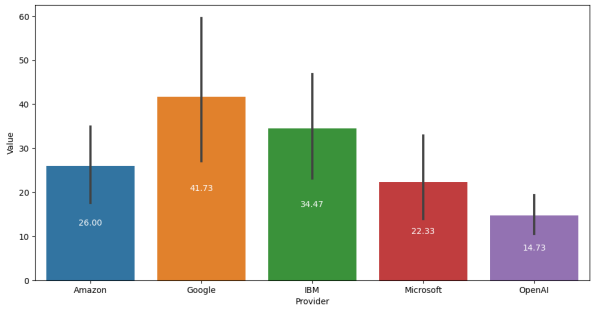


(b) Jaccard Index Box Plot.

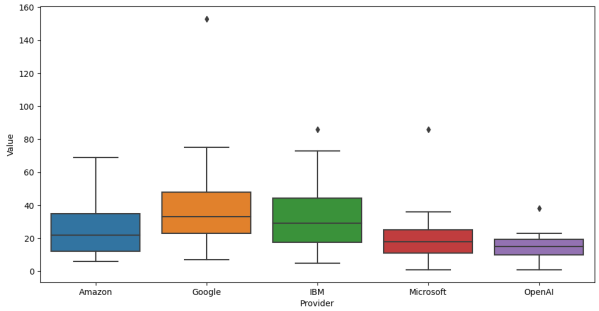


(c) Jaccard Index KDE Plot.

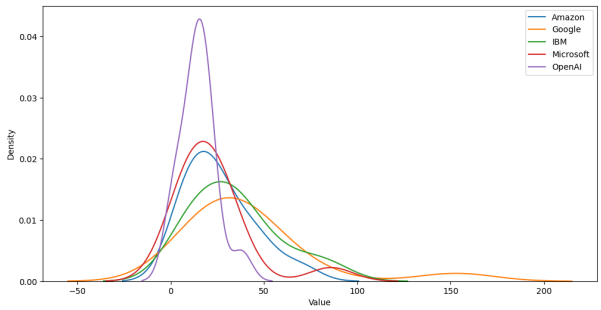
Figure 4: Plots of Jaccard Index Similarity.



(a) Levenshtein Distance Bar Plot.



(b) Levenshtein Distance Box Plot.



(c) Levenshtein Distance KDE Plot.

Figure 5: Plots of Levenshtein Distance Similarity.

services is of great importance, because with their diversity, speech recognition platforms can serve users around the world, allowing people from different countries/cultures to benefit from these technologies. This expands access to information and the opportunity to interact with devices and applications using voice.

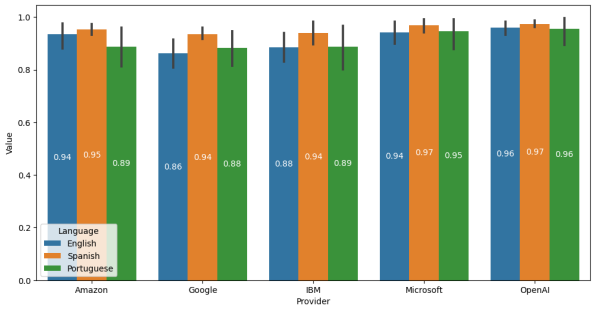
Also generating inclusive communication, as the diversity of languages in speech recognition platforms allows people who speak different languages to communicate more efficiently and inclusively. This is especially important in multicultural environments and in countries with linguistically diverse populations.

Also enabling accessibility for people with hearing impairments or reading difficulties, speech recognition can be a vital tool to facilitate communication and access to information. The availability of multiple languages ensures that people in different regions and cultures can benefit from these technologies to overcome communication barriers.

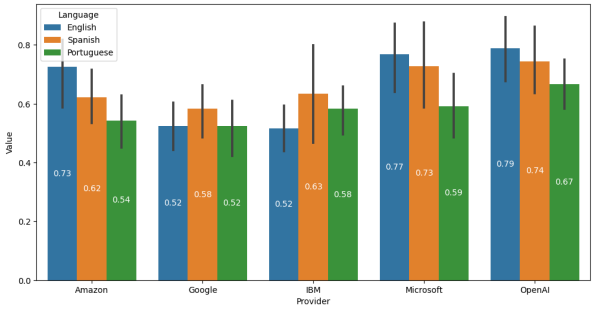
When comparing the Levenshtein distance, where the best values are closest to zero, we can see that OpenAI's STT tool also showed better results compared to its competitors.

3.3. Case Study Threats to Validity

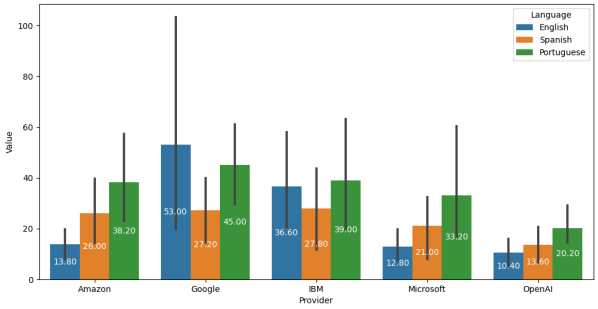
The main threat to the validity of the Case Study arises from the selected sample of video lessons, encompassing 15 videos each with a maximum duration of 30 seconds. Despite this limitation, we employed a semi-random selection strategy to encompass a diverse range of teachers, accents, and languages, thus enhancing the representativity of the educational content provided by DIO in our dataset. This strategy not only mitigates this threat but also serves to reduce selection bias, fostering a more balanced view of the STT providers' performance across varied content. Moreover, by sharing Python scripts in Google Colab, we facilitate future replications and extensions of this study, promoting ongoing research in this field.



(a) Cosine Similarity Bar Plot.



(b) Jaccard Index Bar Plot.



(c) Levenshtein Distance Bar Plot.

Figure 6: Plots of Similarity Tests per Language.

4. Statistical Analysis

Similarity tests indicated that the OpenAI Whisper tool provider was the most advantageous compared to Amazon, Google, IBM, and Microsoft. To strengthen the test results through statistical analysis, a verification process was conducted using a hypothesis test. Two sets of hypotheses were tested. The first set considered the five providers, while the second set considered the five providers and three languages. The hypotheses defined to test the differences among the providers were:

- **Null hypothesis** (H^0): There is no statistically significant difference in the quality of automatic transcripts among different providers, as evaluated by lexical metrics.
- **Alternative hypothesis** (H^1): There is a statistically significant difference in the quality of automatic transcripts among different providers, as measured by lexical metrics.

The first step in selecting a statistical test was to identify the sample of similarity tests (Jaccard, Cosine, Levenshtein) in terms of its normality. These factors were assessed using the Shapiro-Wilk (SW) and Kolmogorov-Smirnov (KS) tests (Morschheuser et al., 2015). Only the Jaccard similarity test was found to be normally distributed ($p = 0.111$ for $\alpha = 0.05$ in the SW test and $p = 0.200$ for $\alpha = 0.05$ in the KS test).

The fact that two of the SW tests indicated that the samples do not follow a normal distribution may suggest that the data are not symmetrically distributed or may contain significant outliers. Consequently, parametric tests rely on specific assumptions, such as data normality, and yield more accurate estimates of population parameters if these assumptions are satisfied.

In some cases, it may be reasonable to apply parametric tests even if the data are not strictly normal, particularly when the sample sizes are large enough that small deviations from the assumptions do not significantly impact the results. However, since this is not the situation with the collected sample, we chose to perform a hypothesis test using the Jaccard similarity test sample. Therefore, One-Way ANOVA tests were conducted to compare the Jaccard sample and determine if there are statistically differences among the providers.

The one-way ANOVA test is a statistical test used to compare the means of three or more independent groups (Ližbetinová et al., 2019). In our case, we have five groups: Group 1 - Amazon, Group 2 - Google, Group 3 - IBM, Group 4 - Microsoft, and Group 5 - OpenAI. The one-way ANOVA test allows us to determine whether significant differences exist between the means of these groups and helps identify which groups are statistically distinct. In this context, all five groups have different means and meet the assumption of normality, which makes it appropriate to investigate whether there are differences among the means.

The ANOVA result indicates a value of 0.002 for $\alpha = 0.05$, suggesting a significant difference among the means of the provider groups. Since the ANOVA test only establishes the presence of a difference among the analyzed groups, further investigation was conducted using the Tukey HSD (Honest Significant Difference) and Bonferroni tests to identify which specific groups exhibit the differences observed in the ANOVA test (Ližbetinová et al., 2019).

The Tukey HSD test is a statistical procedure used to conduct multiple comparisons between the means of different groups. Its purpose is to determine which groups exhibit significant differences in their means. The results of the Tukey HSD test provide information on paired groups for comparison, including confidence intervals and critical values. For each pair of groups, a difference statistic and an associated p-value are provided. Regarding the Bonferroni test, it is a statistical procedure utilized to adjust for significance in multiple comparisons, aiming to control false positive rate. It is employed when multiple null hypotheses are simultaneously tested, mitigating the risk of obtaining significant results by chance.

Bonferroni's method adjusts the significance level by dividing it by the number of comparisons being conducted. Following the execution of multiple comparisons, the p-value of each comparison is compared to the new adjusted significance level (α'). If the p-value is less than α', the null hypothesis for that particular comparison is rejected, indicating a statistically significant difference. Conversely, if the p-value exceeds α', there is insufficient evidence to reject the null hypothesis, leading to the conclusion that no statistically significant difference exists.

For both the Tukey HSD and Bonferroni tests, significant differences were observed among the groups. In the Tukey test, the mean of OpenAI (Group 5) was found to be significantly higher than that of Google (Group 2) and IBM (Group 3), while the mean of Microsoft (Group 4) showed differences compared to Google (Group 2). In the Bonferroni test, OpenAI (Group 5) showed a significant difference compared to Google (Group 2), Microsoft (Group 4) showed a difference compared to Google (Group 2), and IBM (Group 3) showed a difference with OpenAI (Group 5).

Tests of Normality				
Test	**Kolmogorov-Smirnov**		**Shapiro-Wilk**	
	Statistic	**Significance**	**Statistic**	**Significance**
Cosine	0,162	<,001	0,860	<,001
Jaccard	0,057	,200	0,973	,111
Levenshtein	0,175	<,001	0,796	<,001
Interpretation				
Shapiro-wilk: If Significance (p-value) >0.05 is normal.				
Kolmogorov-Smirnov: If Significance >0,05, sample follow the same statistical distribution.				
One-Way ANOVA (One-Way Analysis of Variance)				
Test	**F**		**Significance**	
Jaccard	4,562		0,002	
Interpretation				
If Significance (p-value) <0,05, there are significant differences between at least two groups.				
Tukey HSD (Honest Significant Difference)				
Providers (Groups from 1 to 5)			**Significance**	
OpenAI (Group 5) to Google (Group 2)			0,005	
OpenAI (Group 5) to IBM (Group 3)			0,032	
Microsoft (Group 4) to Google (Group 2			0,037	
Interpretation				
If significance <>0 and <,05, there is a significant difference among the means (groups).				
Bonferroni				
Providers (Groups from 1 to 5)			**Significance**	
OpenAI (Group 5) to Google (Group 2)			0,006	
Microsoft (Group 4) to Google (Group 2			0,047	
IBM (Group 3) to OpenAI (Group 5)			0,041	
Interpretation				
If adjusted significance (α' = α / n), where n is the number of comparisons is <0,05, there is a statistical difference among the groups.				

Table 2: Statistical Tests among Providers.

Such results suggest that the Jaccard result demonstrates OpenAI Provider as the superior choice, as tested in the case study. As a result, we can reject the null hypothesis (H^0) and accept the alternative hypothesis (H^1), confirming a statistically significant difference in the quality of automatic transcripts among the different providers, as measured by the Jaccard lexical metric. Table 2 presents all statistical tests and interpretations for the comparison of providers.

To identify if there is difference in quality of the providers considering the language of the transcriptions, Jaccard sample was also applied to a statistical analysis. The set of hypotheses were:

- **Null hypothesis** (H^0): There is no statistically significant difference in the quality of automatic transcriptions among Portuguese, English, and Spanish, as evaluated by Jaccard lexical metric.
- **Alternative hypothesis** (H^1): There is a statistically significant disparity in the quality of automatic transcriptions for Portuguese among the providers.
- **Alternative hypothesis** (H^2): There is a statistically significant disparity in the quality of automatic transcriptions for English among the providers.
- **Alternative hypothesis** (H^3): There is a statistically significant disparity in the quality of automatic transcriptions for Spanish among the providers.

First, for normality tests (SW and KS test), all the sample considering the groups of languages and the providers (sample of 5 for each group) were normal. The test results are presented in Table 3 along with all the statistical tests.

In this context, a one-way ANOVA test was applied to the sample in its respective language. The results of the ANOVA test for each group are presented in Table 3.

Based on these results, only the alternative hypotheses H^2 show statistically significant differences.

To identify the real difference among the means for each provider, Tukey's HSD (Honestly Significant Difference) and Bonferroni tests were applied. Tukey's HSD test is particularly useful when comparing all possible pairs of groups to identify significant differences, while the Bonferroni test adjusts the significance level to account for multiple comparisons.

Considering the results of multiple comparisons using Tukey's HSD and Bonferroni tests, OpenAI emerges as the superior provider when utilized with the transcriptions in English language.

5. Conclusions

In this paper, we showcased the use of STT as a means to enhance the accessibility of LOs. The *Speech2Learning* Architecture was proposed as a guideline for devising STT-based solutions that further accessible LOs. Our initial PoC, executed in partnership with DIO, provided an essential dataset for the subsequent Case Study. We then analyzed lexical similarity to measure the quality of automatic transcription from the world's leading providers: Amazon, Google, IBM, Microsoft, and OpenAI. Finally, our Statistical Analysis of the quantitative Case Study data revealed significant differences in terms of language and provider.

Tests of Normality				
Test	**Kolmogorov-Smirnov**		**Shapiro-Wilk**	
	Statistic	**Significance**	**Statistic**	**Significance**
PT-BR	0,121	,200	0,961	0,427
EN	0,114	,200	0,942	0,169
ES	0,951	<,001	0,951	0,264
One-Way ANOVA - Brazilian Portuguese				
Test	**F**		**Significance**	
PT-BR	1,058		0,403	
Tukey HSD (Honest Significant Difference)				
Providers (Groups from 1 to 5)			**Significance**	
OpenAI (Group 5) to Google (Group 2)			0,359	
Bonferroni				
Providers (Groups from 1 to 5)			**Significance**	
OpenAI (Group 5) to Google (Group 2)			0,748	
One-Way ANOVA - English				
Test	**F**		**Significance**	
EN	4,88		0,007	
Tukey HSD (Honest Significant Difference)				
Providers (Groups from 1 to 5)			**Significance**	
OpenAI (Group 5) to Google (Group 2)			0,041	
OpenAI (Group 5) to IBM (Group 3)			0,034	
Bonferroni				
Providers (Groups from 1 to 5)			**Significance**	
IMB (Group 3) to OpenAI (Group 5)			0,047	
One-Way ANOVA - Spanish				
Test	**F**		**Significance**	
ES	0,931		0,466	
Tukey HSD (Honest Significant Difference)				
Providers (Groups from 1 to 5)			**Significance**	
OpenAI (Group 5) to Google (Group 2)			0,539	
Bonferroni				
Providers (Groups from 1 to 5)			**Significance**	
All groups			1	

Table 3: Statistical Tests among Languages.

Considering the statistical tests conducted to compare the differences observed between the means of the provider groups in the Case Study, the results, based on the Jaccard Index metric, confirmed the presence of statistically significant differences among the averages. These statistical validations add empirical weight to the significance of choosing an appropriate STT provider for educational applications. The Tukey HSD test revealed a significant difference for the OpenAI provider when compared to Google and IBM, while also indicating a significant difference for the Microsoft provider in comparison to Google. These findings support the claim that OpenAI and Microsoft providers offer the highest quality automatic transcription services.

Additionally, when analyzing the statistical results considering the different provider groups and the languages of the transcriptions (Portuguese, English, and Spanish), the one-way ANOVA test revealed a significant difference only among the providers generating transcriptions in English. More specifically, there was a significant difference between OpenAI and Google, as well as between OpenAI and IBM. In other words, the results indicate that, despite the initial findings across all providers, the differences in means are particularly prominent when considering transcriptions in English. This suggests that providers still need improvements specifically related to other languages, such as Portuguese and Spanish. These findings also open up opportunities for conducting new case studies and statistical tests involving different languages.

Going further, the specificities of each selected language in our study can be investigated by considering their origins and how they are treated by each provider. Although all three languages have their origins and classification as Romance languages, derived from Latin, the results show significant differences in providers' behaviors in generating transcripts for the English language. This suggests a potential bias towards English in STT algorithms, which warrants further investigation to ensure truly inclusive educational technologies.

We acknowledge that our study has limitations, including the constrained dataset of 15 videos that may not capture the full diversity of linguistic features and accents. Moreover, focusing on the Jaccard Index as the main metric for lexical analysis might not capture the nuanced variations in transcription quality. Moving forward, we aim to expand our dataset to include a wider range of languages, dialects, and educational topics, addressing the reviewer's concern for a more universal application. Additionally, exploring alternative metrics and statistical tests will be crucial for a more comprehensive understanding.

In light of our findings and feedback from industry partners, we plan to explore further possibilities for the *Speech2Learning* Architecture by promoting new language variants, thereby improving its applicability and inclusiveness. In this sense, we intend to expand the use of our architecture to the context of text-based sign language avatars, enabling the creation of assistive technologies for the deaf community. This expansion could diversify our dataset and foster new research in spoken and sign languages.

Acknowledgment

The authors would like to thank the Brazilian funding agencies – São Paulo Research Foundation (FAPESP) under grant #2018/26636-2; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; and CNPq. We extend our gratitude to EdTech DIO for the partnership on both the PoC and the Case Study, enabling market insights through one of the largest e-learning platforms in Latin America.

References

Bosker, H. R. (2021). Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies. *Behavior Research Methods*.

FalvoJr, V., Scatalon, L. P., & Barbosa, E. F. (2020). The Role of Technology to Teaching and Learning Sign Languages: A Systematic Mapping. *Proceedings of the 50th Annual Frontiers in Education Conference (FIE)*.

Fleischmann, A. C., Cardon, P., & Aritz, J. (2021). Acceptance of speech-to-text technology: Exploring language proficiency and psychological safety in global virtual teams. *Proceedings of the 54th HICSS*.

Gartner. (2023). *Magic quadrant for cloud ai developer services*. https://bit.ly/3Z0ZNQn

Hapke, H., Howard, C., & Lane, H. (2019). *Natural language processing in action: Understanding, analyzing, and generating text with python*. Simon; Schuster.

Hasan, H. M. M., Islam, M. A., Hasan, M. T., Hasan, M. A., Rumman, S. I., & Shakib, M. N. (2020). A spell-checker integrated machine learning based solution for speech to text conversion. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*.

Homburg, D., Thieme, M. S., Völker, J., & Stock, R. (2019). RoboTalk - prototyping a humanoid robot as speech-to-sign language translator. *Proceedings of the 52nd HICSS*.

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.

Kumar, L. A., & Renuka, D. K. (2023). *Deep learning approach for natural language processing, speech, and computer vision: Techniques and use cases*. CRC Press.

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. *4th International Conference on Cyber and IT Service Management*.

Ližbetinová, L., Štarchoň, P., Weberová, D., Nedeliaková, E., & Juříková, M. (2019). The approach of smes to using the customer databases and crm: Empirical study in the slovak republic. *Sustainability*, *12*(1).

Majumdar, D. (2022). *Text analysis and distant reading using r* (2022.09.13). https://slcladal.github.io/lexsim.html

Manalu, D. R., Rajagukguk, E., Siringoringo, R., Siahaan, D. K., & Sihombing, P. (2019). The development of document similarity detector by jaccard formulation. *International Conference of Computer Science and Information Technology (ICoSNIKOM)*.

Martin, R. C. (2017). *Clean architecture: A craftsman's guide to software structure and design*. Prentice Hall.

Martin, R. C. (2021). *Clean craftsmanship: Disciplines, standards, and ethics*. Prentice Hall.

Mayer, R. E., & Fiorella, L. (2021). *The cambridge handbook of multimedia learning* (3rd ed.). Cambridge University Press.

Mohana, H., & Suriakala, M. (2018). Integrated cosine and tuned cosine similarity measure to alleviate data sparsity issues for personalized recommendation. *3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, 41–49.

Morschheuser, B., Henzi, C., & Alt, R. (2015). Increasing intranet usage through gamification – insights from an experiment in the banking industry. *Proceedings of the 48th HICSS*.

Noel, S. (2020). Human computer interaction(hci) based smart voice email (vmail) application - assistant for visually impaired users (viu). *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 895–900.

Parakh, A., Subramaniam, M., & Chundi, P. (2022). A framework for incorporating serious games into learning object repositories through experiential learning. *Proceedings of the 55th HICSS*.

Ristanti, P. Y., Wibawa, A. P., & Pujianto, U. (2019). Cosine similarity for title and abstract of economic journal classification. *5th International Conference on Science in Information Technology (ICSITech)*.

Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Softw. Engg.*, *14*(2), 131–164.

Sasu, D. (2019). *Developing a functional natural language processing system for the twi language with limited data* [Doctoral dissertation].

Sommerville, I. (2015). *Software engineering* (10th ed.). Pearson.

Sugiarto, Diyasa, I. G. S. M., & Diana, I. N. (2020). Levenshtein distance algorithm analysis on enrollment and disposition of letters application. *6th Information Technology International Seminar (ITIS)*, 198–202.

Sulaiman, N. H., & Mohamad, D. (2012). A jaccard-based similarity measure for soft sets. *IEEE Symposium on Humanities, Science and Engineering Research*, 659–663.

Tanberk, S., Dağlı, V., & Gürkan, M. K. (2021). Deep learning for videoconferencing: A brief examination of speech to text and speech synthesis. *6th International Conference on Computer Science and Engineering (UBMK)*, 506–511. https://doi.org/10.1109/UBMK52708.2021.9558954

Zhang, S., Hu, Y., & Bian, G. (2017). Research on string similarity algorithm based on levenshtein distance. *IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2247–2251.