Mauro Ricardo da Silva Teófilo

Enabling Deaf or Hard of Hearing Accessibility in Live Theaters through Virtual Reality

Manaus, Brazil

2019

Mauro Ricardo da Silva Teófilo

Enabling Deaf or Hard of Hearing Accessibility in Live Theaters through Virtual Reality

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Computing of the Universidade Federal do Amazonas.

Universidade Federal do Amazonas - UFAM

Programa de Pós-Graduação em Informática

Supervisor: Prof. Dr. -Ing. Vicente Ferreira de Lucena Junior

Manaus, Brazil

2019

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Teófilo, Mauro Ricardo da Silva
T356e Enabling Deaf or Hard of Hearing Accessibility in Live Theaters through Virtual Reality / Mauro Ricardo da Silva Teófilo. 2019
105 f.: il. color; 31 cm.

Orientador: Vicente Ferreira de Lucena Junior
Tese (Doutorado em Informática) - Universidade Federal do Amazonas.

1. Accessibility. 2. Virtual Reality. 3. Semantic Similarity. 4. Human Computer Interaction. I. Lucena Junior, Vicente Ferreira de II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Enabling Deaf or Hard of Hearing Accessibility in Live Theaters Through Virtual Reality"

MAURO RICARDO DA SILVA TEÓFILO

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Vicente Ferreira de Lucena Junior - PRESIDENTE

Prof. Raimundo da Silva Barreto - MEMBRO INTERNO

Prof. Waldir Sabino da Silva Júnior - MEMBRO EXTERNO

Prof. Eduardo Lázaro Martins Naves - MEMBRO EXTERNO

Prof. Paulo Henrique da Fonseca Melo - MEMBRO EXTERNO

Manaus, 12 de Abril de 2019

For my mother.

Acknowledgements

The realization of this work was only possible due to the several people's collaboration, to which desire to express my gratefulness.

My forever interested, encouraging and always enthusiastic mother Rocicler Teófilo: she was always keen to know what I was doing and how I was proceeding, although it is likely that she has never grasped what it was all about!

I bow in ovation to my wife and my mother in law for their care and kindness.

I would like to thank from a special way to Professor Vicente F. Jucena Jr., thesis adviser, for his guidance and support throughout this study and specially for his confidence in me.

A very special gratitude goes out to all down at Sidia and Samsung for helping and providing the funding for the work.

With a special mention to Juliana Postal, Alvaro Lourenço, Francimar Maciel, Victor Santos, Maurílio da Silva, Luis Rojas, Josiane Nascimento, Alvaro Gonçalves, Julie Neviere, the develop and research team. It was fantastic to have the opportunity to work majority of my research with their support.

Finally, I would like to thank to the Almighty God, thank you for the guidance, strength, power of mind, defense, skills and for giving us a healthy life.

“Tem que ter as dificuldade,
pra se vencer. "
(Mestre Gabriel)

Abstract

Recent advancements in Virtual Reality (VR) have made them a potential technology to improve understanding between Deaf or Hard of Hearing (DHH) and hearing people. Based on that, this study aims to extend these advances to enable live entertainment like theater plays to DHH people.

At first, this work presents a survey, which covers some findings of the accessibility research using virtual and augmented reality systems, ranging years from 1996 to nowadays, and fields such as children, autism, motor rehabilitation, Parkinson disease, and inclusion for the impaired.

After this initial review, it's presented one solution in details, which is a live theater accessibility service for the deaf and hard of hearing people, aiming to be a concrete solution designed to bringing accessibility using virtual reality technology combined with Automatic Speech Recognition (ASR), Sentence Prediction and Speech Correction to generate text and sign language subtitling. In order to evaluate this method efficiency, a quantitative and qualitative study were performed and results showed that DHH spectators had good understanding of all evaluated theater plays and also good satisfaction using the proposed method. The best results are related to text subtitling. Regarding sign language subtitling, it is a promising technology, but a huge effort is necessary to start a standard for displaying definition in virtual or augmented reality device.

Other main contribution of this work is to present the procedures and its results executed to evaluate the integration of ASR and Speech Correction based on Semantic Similarity of a solution designed to bring accessibility to DHH People in live theaters. Five datasets were submitted to two different ASRs and its outputs to the module of Speech Correction considering three groups of pairs to observe the semantic and syntactic errors presented by the modules as well as its performance under different configurations representative of possible scenarios of actual theatrical plays. The ASRs presented an error of 50% in average when applied to the audio of an actual play. The module for semantic similarity presented an average error of 18% on sentences not modified by the ASR. However, the output of the Semantic Similarity module is affected by the error introduced by the ASRs. Speech Correction based on Semantic Similarity would present less error than Syntactically based.

Keywords: Accessibility. Virtual Reality. Semantic Similarity. Deaf of Hard of Hearing.

List of Figures

List of Tables

List of abbreviations and acronyms

VR	Virtual Reality
AR	Augmented Reality
VR	Virtual Environment
DHH	Deaf or Hard Hearing
ASR	Automatic Speech Recognition
SP	Sentence Prediction
SC	Speech Correction
STT	Speech to Text
NLP	Natural Language Processing
HCI	Human Computer Interaction
ASD	Autism Spectrum Disorders
PWDS	Persons With Disabilities
STS	Semantic Textual Similarity
RWC	Rear Window Captioning
WRR	Word Recognition Rate
xRT	Realtime Factor
ASSIN	Avaliação de Similaridade Semântica e Inferência Textual
API	Application Programming Interface
REST	Representational State Transfer
gRPC	open source Remote Procedure Call
CBOW	Continuous Bag-of-Words model
VSM	Vector Space Models

Contents

1 Introduction

The main objective of this chapter is to present an introduction to this thesis coming up with the following explanation: *a)* the general current context regarding accessibility and Virtual Reality technology applied as an assistive tool, *b)* the thesis motivation, where it is indicated the benefits of this thesis proposal, *c)* the proper definition of the addressed problem, *d* the thesis objectives list, the general and specific objectives are presented, *e)* the brief summary of the proposed method, which aims to solve the addressed problem and reach the listed objectives, and *f)* the thesis structure that is divided in three huge parts to better organize it, which are: literature overview, the method, and analysis.

1.1 Context

According to the World Health Organization, 15% of the entire world population lives with some type of disability (8). This indicator justifies the amplification of works on Assistive Technology over the last years also inferring on the needs of evaluative methods and reliable data.

The emerging hot topic of Virtual Reality (VR) and Augmented Reality (AR) comes bundled with at least two decades of relevant research threads on the accessibility field, leveraging actionable legacy findings on the cognitive and physical rehabilitation.

With the ever growing availability of VR/AR technologies, and with several authors now inspecting positive results out of studies that so far helped people with learning disabilities, relieved patients from anxiety and phobias, improved social participation of the visual and auditory impaired, restored mobility for patients with spinal cord injury, and postural balance for those in post-stroke and Parkinson disease, VR/AR is now on the verge of shaping new standards for a broad spectrum of the accessibility field.

In this context, a rising number of researches around virtual environments has been scoping its possible benefits to compensate people for hearing disadvantages by improving their autonomy.

1.2 Motivation

Our society was designed from the ground up to accommodate the needs of able-bodied individuals, so there are times when everyday situations can become a struggle for those who are deaf. There are barriers to basic access that limit the rights and freedoms of those who can not hear, subtly perpetuating an existing structure of oppression.

People who are deaf deserve access to every moment of shared collective joy, pain, awe, introspection, and outward rage that can be elicited through performance. To deny a person this experience is to deny them access to the very culture in which they live and the possibility of meaningful human connection. From live theater shows, to the national anthem at a football game, to a pop concert, and everything in between, each and every person in attendance should be able to share it.

Theaters are a place of public accommodation where people from all walks of life are entitled to share an experience. A culturally significant play such as "Death of a Salesman" by Arthur Miller, can be meaningful to people's lives. Everyone deserves to share in that excitement and collective social moment if they so choose.

Performing arts are an outlet for self expression; a way to explore complex human emotions and taboo topics. Theaters, venues, and even musicians themselves are being pressured to evolve to meet the demands of culturally aware audiences, who value inclusion.

1.3 Problem Definition

Even with all current technology, progressive innovators have failed to address some of the real problems that persist when it comes to accessibility. The problem to be solved by the proposed solution is to bring accessibility in live theaters to Deaf of Hard of Hearing (DHH) people.

There are a couple of attempts of bringing accessibility for DHH to movie theaters, which is a very similar environment. Although, live theaters are a more complex scenario due actors timing and improvisation.

Based on literature review, movie theaters are still not friendly places for deaf. Even using current technologies, like closed captioning devices that deaf movie-goers are given to use make it difficult to focus on both the film, which is in the background, and the screen, which is in the foreground. The constant shift in focus can be exhausting, and can also cause the viewer to miss a great deal of the action in the movie. Then, this kind of solution will not be used to solve the raised problem.

It becomes clear that open captioning (with the transcript right on the screen) is the preferred accommodation for deaf audiences. Thus, VR technology can be helpful to emulates this scenario in a live theaters context.

Other problem is related to part of DHH community which don't know to read, they just use sign language to communicate with each other. The method proposed in this thesis must provide solution for this cited deaf population.

All intrinsic characteristic of live theater must be taken in consideration in proposed method, like play improvisation, changing in speeches, substitution of some words of speech,

removal or addition of speeches, etc.

There is some deaf, which can lipread if sitting close to the stage. However, the closer you sit, the more expensive ticket prices become, so lip-reading from a distance is almost impossible. This proposed solution mustn't be location based to avoid expensive ticket for deaf population.

In summary, the problem to be solved is to enable DHH people to access a live theater.

1.4 Objectives

1.4.1 General Objective

The main target of this thesis is to create and evaluate a method to enable Deaf of Hard of Hearing (DHH) people to access live theaters, meeting the needs of DHH audiences.

1.4.2 Specific Objectives

1. Create a text subtitling method to enable DHH people to follow a play using a VR device.
2. Create a sign language subtitling method to enable DHH people to follow a play.
3. Use automatic speech recognition and sentence prediction to show subtitling in real time based on script play.
4. Use speech correction to allow actors speech improvisation during the play.
5. Evaluate the proposed solution in a real scenario. The evaluation must be done with target audience (DHH people).

1.5 Thesis Contribution

This thesis aims to contribute with accessibility topic proposing a method to enable the DHH people to attend a live theater play. Thus, this thesis proposes a method using VR technology combined with ASR, sentence prediction, speech correction to generate text and sign language subtitling in real time. Then, this method will allow to include DHH people socially, because it gives the opportunity for them to attend in live events.

The sentence prediction technology will ensure the behavior of retrieving subtitles in realtime while scene performance is occurring. The play script will be the basis for the sentence prediction. Even play script usage, in live theaters, actors actors improvisation are very usual. Then, it is expected that in most of cases the actor speaks a sentence that is semantically similar with the corresponding sentence in the play script. Is for that reason that a suitable approach for Speech Correction is Semantic-Similarity-based sentence matching.

Aiming to allow DHH people to understand a play, a subtitle is presented during a play performance powered by a VR device. The subtitle can be a text or a sign language. This subtitle is generated when an actor in a play speaks, the voice is captured by microphones, then this voice information is sent to the ASR module to transform the voice into text. The ASR's outputted text is corrected by a module of speech correction that predates the play script for a semantically similar sentence with that spoken by the actor. The output of the module of Speech Correction is then sent as a subtitle to all VR devices used by DHH spectators.

1.6 Thesis Structure

The first part of this thesis, part I, is composed by a literature review, which is composed by two chapters, where in Chapter 2, Background, a brief concept of main technologies used in proposed solution is summed up and in Chapter 3, Related Work, a survey about state of art related to accessibility field improvements through Virtual Reality technology is summarized.

In the second part, part II, The Proposed Method Overview, the concept of proposed method to enable Deaf of Hard of Hearing (DHH) accessibility in live theaters through virtual reality is presented in Chapter 4, Subtitling Method to DHH People in Live Theaters, where the cited method to solve the proposed problem is described, as well as, the text and sign language subtitling are presented in details.

Analysis, part III, which is the third part of this thesis and it covers all experiments and results obtained during this thesis development, including the chapter 5, Experiments and Results related to ASR, where is presented a comparison between the ASRs used in proposed method based on four metrics. In the chapter 6, Experiments and Results related to Speech Correction through Semantic Similarity, the Semantic Similarity used for speech correction is evaluated in details. The chapter 7, Experiments and Results related to Subtitling Method, summarize all finding related to subtitling method tested by target audience (DHH community). Then, in Chapter 8, Conclusion, all main findings are listed and summarized.

Part I

Literature Overview

2 Background

This chapter provides a brief overview of some computational techniques that have been used during this thesis development. Technologies as Automatic Speech Recognition (ASR), Sentence Prediction (SP), Semantic Textual Similarity (STS), which was used to Speech Correction (SC), where is highlighted the Word Embedding by Word2vec technique. All these technologies were used to concept and to implement the proposed solution to be evaluated in this thesis.

2.1 Automatic Speech Recognition

Automatic Speech recognition or ASR is a sub-field of natural language processing that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers (9). It is also known as *computer speech recognition*, or just *speech to text* (STT). The goal of ASR is to build systems that map from acoustic signals to a string of words. It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields. Some speech recognition systems require *training"* (also called *enrollment*) where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called *speaker independent* systems. Systems that use training are called "speaker dependent" (10). Speech recognition applications include voice user interfaces such as voice dialing, call routing, domotic appliance control, search, simple data entry, preparation of structured documents, speech-to-text processing, and aircraft.

Speech recognition problems has some important points to be considered, the first one is the vocabulary size. Speech recognition is easier if the number of distinct needed words to recognize is smaller. So tasks with a two word vocabulary, like yes versus no detection, or an eleven word vocabulary, like recognizing sequences of digits (11), in what is called the digits task, are relatively easy. On the other end, tasks with large vocabularies, like transcribing human-human telephone conversations, or transcribing broadcast news, tasks with vocabularies of 64,000 words or more, are much harder. Second is isolated word recognition, in which each word is surrounded by some sort of pause, that is much easier than recognizing continuous speech, in which words run into each other and have to be segmented (12). Continuous speech tasks themselves vary greatly in difficulty. For example, human-to-machine speech turns out to be far easier to recognize than human-to-human speech. That is, recognizing speech of humans talking to machines, either reading out loud in read speech (which simulates the dictation task), or conversing with speech dialogue systems, is relatively easy. Recognizing the speech of two

humans talking to each other, in conversational speech recognition, for example for transcribing a business meeting or a telephone conversation, is much harder. It seems that when humans talk to machines, they simplify their speech quite a bit, talking more slowly and more clearly.

The third is channel and noise. Commercial dictation systems, and much laboratory research in speech recognition, is done with high quality, head mounted microphones (13). Head mounted microphones eliminate the distortion that occurs in a table microphone as the speakers head moves around. Noise of any kind also makes recognition harder. Thus recognizing a speaker dictating in a quiet office is much easier than recognizing a speaker dictating in a noisy car on the highway with the window open, and finally accent or speaker-class characteristics. Speech is easier to recognize if the speaker is speaking a standard dialect, or in general one that matches the data the system was trained on (14). Recognition is thus harder on foreign accented speech, or speech of children (unless the system was specifically trained on exactly these kinds of speech).

In noisy channel model theory, it is assumed that a sentence to be predicted is the result of a passage through a noisy channel (15). So if it's found a representation of this noisy channel, we can test all the possible sentences of a language in this model (decoder) and see which result makes the best match.

In the method proposed in this thesis detailed in chapter 3.7 is used two ASRs, Google Cloud Speech-to-Text and CPqD ASR. They are described in the next sections.

2.1.1 Google Cloud Speech-to-Text

Google Cloud Speech-to-Text enables developers to convert audio to text by applying powerful neural network models in an easy-to-use API. The API recognizes 120 languages and variants to support your global user base. You can enable voice command-and-control, transcribe audio from call centers, and more. It can process real-time streaming or prerecorded audio, using Google's machine learning technology (16).

A Speech-to-Text API synchronous recognition request is the simplest method for performing recognition on speech audio data. Speech-to-Text can process up to 1 minute of speech audio data sent in a synchronous request. After Speech-to-Text processes and recognizes all of the audio, it returns a response.

Speech-to-Text has three main methods to perform speech recognition. These are listed below:

- Synchronous Recognition (REST and gRPC) sends audio data to the Speech-to-Text API, performs recognition on that data, and returns results after all audio has been processed. Synchronous recognition requests are limited to audio data of 1 minute or less in duration.
- Asynchronous Recognition (REST and gRPC) sends audio data to the Speech-to-Text API and initiates a Long Running Operation. Using this operation, you can periodically poll

for recognition results. Use asynchronous requests for audio data of any duration up to 180 minutes.

- Streaming Recognition (gRPC only) performs recognition on audio data provided within a gRPC bi-directional stream. Streaming requests are designed for real-time recognition purposes, such as capturing live audio from a microphone. Streaming recognition provides interim results while audio is being captured, allowing result to appear, for example, while a user is still speaking.

For the method proposed in this thesis, the streaming STT API needs to be used to allow continuous translation during a theater play section. A streaming Speech-to-Text API recognition is designed for real-time capture and recognition of audio, within a bi-directional stream. The application can send audio on the request stream, and receive interim and final recognition results on the response stream in real time. Interim results represent the current recognition result for a section of audio, while the final recognition result represents the last, best guess for that section of audio.

2.1.2 CPqD Automatic Speech Recognition

CPqD ASR was built following the client/server paradigm as illustrated in Figure 1, with the recognition engine integrated with a server that offers the proprietary WebSocket and REST interfaces, which it allows client applications to have access to speech recognition capabilities . An independent service can also be installed to provide the standardized Media Resource Control Protocol (MRCP) interface, commonly used in integration with IVR (Interactive Voice Response).

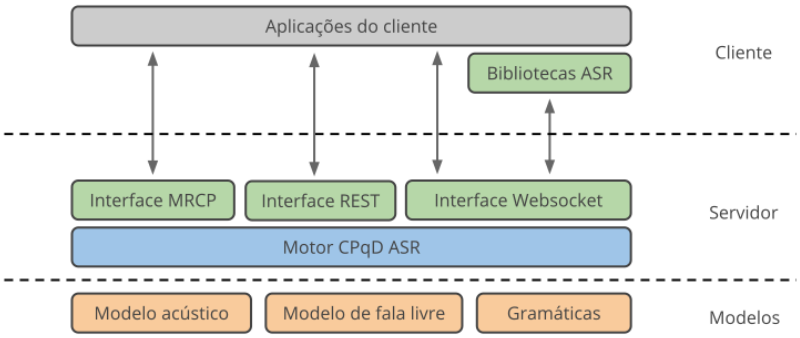

Figure 1 – CPqD ASR Components (1).

The acoustic model and the free speech models are installed through independent installation packages (language packs), which must conform to the characteristics that one wishes to attend. For example, language packs used in an installation for Brazilian Portuguese are different

from those used in a Spanish installation. Language packs designed to work with 16 kHz sampled audio should not be used in applications involving telephone calls (with an audio sampling rate of 8 kHz).

Another important factor involving language packs is the "context" in which speech recognition will work. Some packages may be more specific to some contexts, privileging a particular area (medical area, sports area, etc.). Other packages can be more general, covering different contexts. Generally, stricter packages will present greater accuracy when applied to the correct contexts.

Some predefined grammars, called built-in grammars, are distributed with CPqD ASR, and can be installed and used in recognition. Specific grammars can also be created by the application developer himself, with the help of tools offered by CPqD ASR. To facilitate the development of applications, some client libraries are available that simplify the integration process with CPqD ASR.

2.2 Sentence Prediction

Sentence Prediction in natural language processing is the problem of guessing which sequence of words is likely to continue a given initial text fragment. Sentence prediction techniques are well-established methods in the field of AAC (Augmentative and Alternative Communication) that are frequently used as communication aids for people with disabilities, accelerate the writing; reduce the effort needed to type; suggest the correct word (17). Most common implementation is by Ngrams, a statistical language modeling approach. Ngram is a way to find the language model from a text corpus that is basically a probability distribution $P(s)$ over all possible sentences s of this corpus (18). Probabilities are about to observe occurrences and count things, but we need to know beforehand what we want to observe and count (19), in natural language processing we count words that come from corpora, which are collections of text or speech, a well-known corpora is the Brown corpus(singular of corpora) which has 1 million words in English language from many sources(novels, academic, newspaper), but it is also common to see researches build their own corpus. When we deal with text as our source data for learning we usually use strategies of preprocessing to normalize the text, filter noise or enhance features, some of these strategies are text tokenization, stopwords removal, stemming and lemmatization (20).

Tokenization of text is the task of segnormalization menting running text into words, and normalization, the task of putting words/tokens in a standard format, one commonly used tokenization standard is known as the Penn Treebank tokenization standard, used for the parsed corpora released by the Linguistic Data Consortium (LDC), the source of many useful datasets (21). Stopwords are a set of words that does not contribute significantly to sentences semantics, useful only to compose sentence's structure, punctuation and words belonging to classes like articles, interjections, auxiliary verbs, particles, pronouns and some others are often considered

stopwords, despite not being so useful to problems like sentiment analysis or opinion mining, these stopwords and punctuation can still be important to develop grammatical parsers for corpora (22). Stemming and lemmatization both do reduction of inflected words, inflections modifies words to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, and mood, in stemming, depending on algorithm, different rules are applied to reduce the words to a root form which not necessarily matches its morphological root form, as it is done in lemmatization (23). The complete process of text normalization is better illustrated in Figure 2.

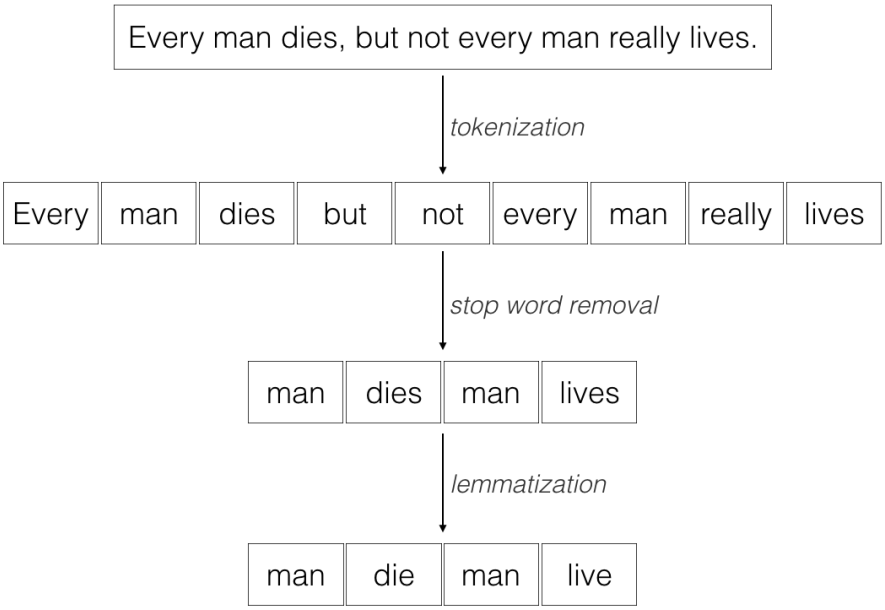

Figure 2 – Text normalization steps.

2.3 Semantic Textual Similarity

In Natural Language Processing (NLP), semantic similarity plays an important role and one of the fundamental tasks for many NLP applications and its related areas. Semantic Textual Similarity (STS) can be defined by a metric over a set of documents with the idea is to finding the semantic similarity between them. Similarity between the documents is based on the direct and indirect relationships among them (24, 25). These relationships can be measured and recognized by the presence of semantic relations among them. Identification of STS in short texts was proposed in 2006 in the works reported in (26, 27). After that, focus was shifted on large documents or individual words.

After that, since 2012 the task of semantic similarity is not only limited to find out the similarity between two texts, but also to generate a similarity score from 0 to 5 by different SemEval tasks[1]. In this task, a scale of 0 means unrelated and 5 means complete semantically

[1] http://ixa2.si.ehu.es/stswiki/index.php

equivalence.

Since its inception, the problem has seen a large number of solutions in a relatively small amount of time. The central idea behind the most solution is that, the identification and alignment of semantically similar or related words across the two sentences and the aggregation of these similarities to generate an overall similarity (28).

Measuring semantic similarity between texts can be categorized into the following ways: (i) topological (ii) statistical similarity (iii) semantic based (iv) vector space model (v) word alignment based and (vi) machine learning. Among these methods, topological studies plays an important role to understand intended meaning of an ambiguous word, which is very difficult to process computationally. For many NLP related task it is important to understand the semantic relation between the word/ concepts. To decompose such systems we need to work with word level relation and those can be considered as hierarchical, associative and equivalence (29).

2.3.1 Word Embedding

Distributed representations for words were proposed and have become extremely successful (30). The main advantage is that the representations of similar words are close in the vector space, which makes generalization to novel patterns easier and model estimation more robust (31).

Word embedding is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.

Vector space models (VSMs) represent (embed) words in a continuous vector space where semantically similar words are mapped to nearby points ('are embedded nearby each other'). VSMs have a long, rich history in NLP, but all methods depend in some way or another on the Distributional Hypothesis, which states that words that appear in the same contexts share semantic meaning. The different approaches that leverage this principle can be divided into two categories: count-based methods (e.g. Latent Semantic Analysis), and predictive methods (e.g. neural probabilistic language models).

Count-based methods compute the statistics of how often some word co-occurs with its neighbor words in a large text corpus, and then map these count-statistics down to a small, dense vector for each word. Predictive models directly try to predict a word from its neighbors in terms of learned small, dense embedding vectors (considered parameters of the model).

Word2vec is a particularly computationally-efficient predictive model for learning word embeddings from raw text. It comes in two flavors, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model models recently proposed by (32). These models learn word representations using a simple neural network architecture that aims to predict the neighbors of a word like shown in Figure 3. Algorithmically, these models are similar, except that CBOW predicts

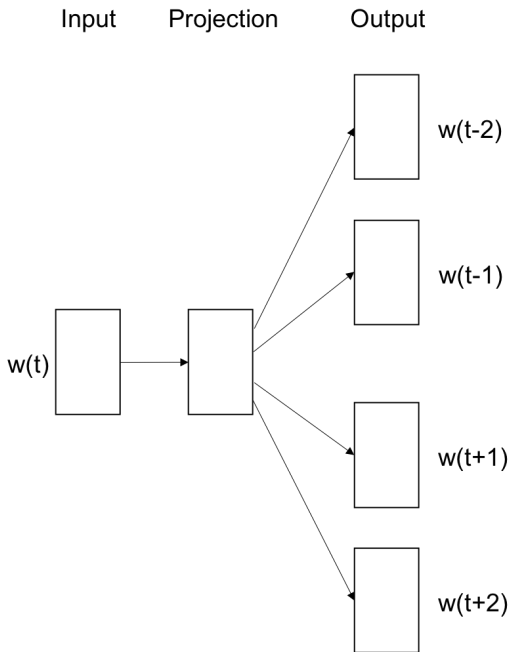

Figure 3 – On Skip-Gram a given word is used to predict its neighbors based on context (2).

target words (e.g. *mat*) from source context words (*the cat sits on the*), while the skip-gram does the inverse and predicts source context-words from the target words. This inversion might seem like an arbitrary choice, but statistically it has the effect that CBOW smoothes over a lot of the distributional information (by treating an entire context as one observation). For the most part, this turns out to be a useful thing for smaller datasets. However, skip-gram treats each context-target pair as a new observation, and this tends to do better when we have larger datasets.

According to Mikolov et al., in (33) distributed representations of words capture surprisingly many linguistic regularities, and that there are many types of similarities among words that can be expressed as linear translations. For example, vector operations *king* - *man* + *woman* results in a vector that is close to *queen*.

To better exemplify a vector representation of words, after training has finished in Figure 4 can be visualized the learned embeddings using t-SNE[2]. As expected, words that are similar end up clustering nearby each other.

2.3.2 Universal Dependencies

Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of (universal) Stanford dependencies (34), Google universal part-of-speech tags (35), and the Interset interlingua for morphosyntactic

[2] https://www.tensorflow.org/tutorials/representation/word2vec

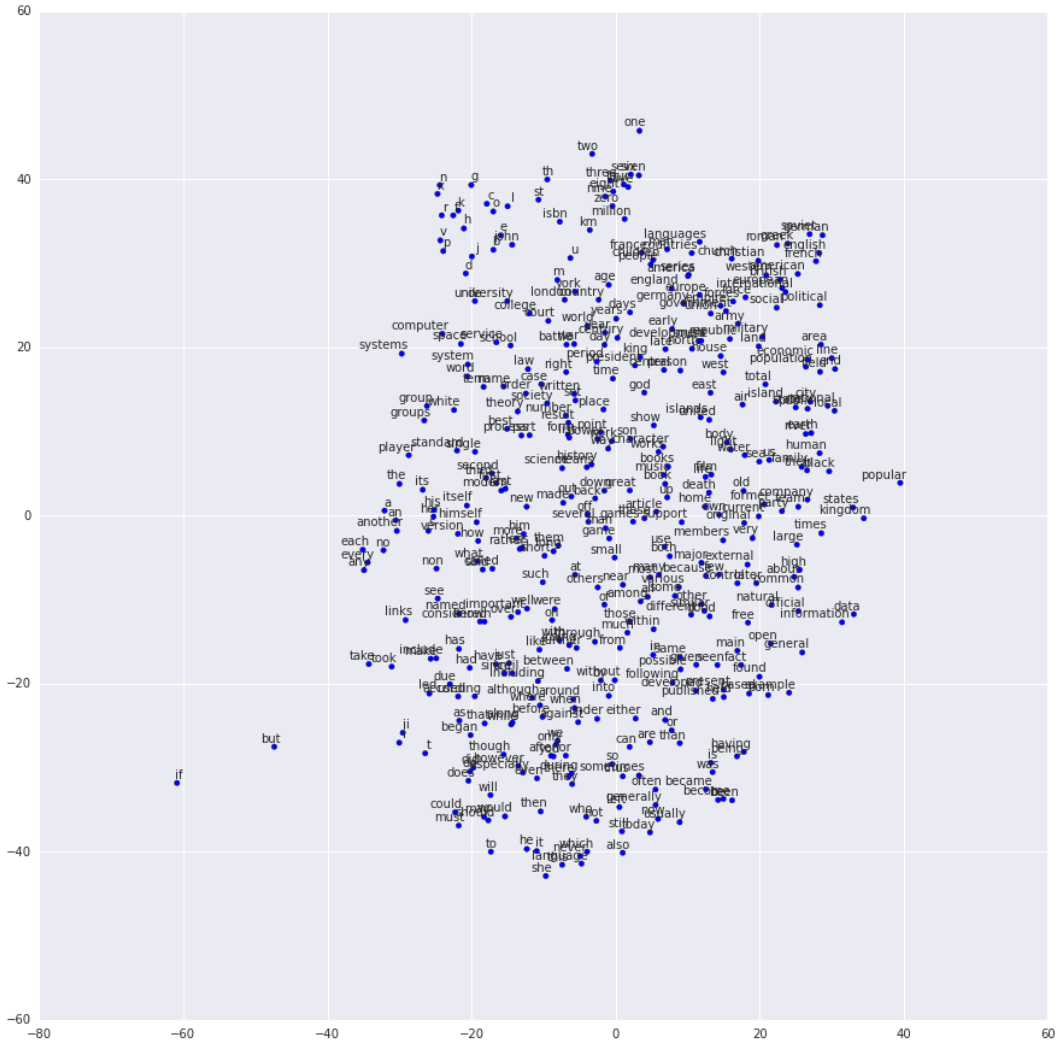

Figure 4 – Learned embeddings using t-SNE (3).

tagsets (36). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary.

This is illustrated in the Figure 5 where is showed parallel examples from English, Bulgarian, Czech and Swedish, where the main grammatical relations involving a passive verb, a nominal subject and an oblique agent are the same, but where the concrete grammatical realization varies.

The module for Semantic Similarity evaluated in this thesis is implemented through the python framework for natural language processing *Spacy*[3]. It integrates an statistical model that follows the architecture of a Word Embeddings by Word2Vec trained using the Universal Dependencies[4](37) and WikiNER[5](38) Corpuses.

[3] https://spacy.io/ (accessed February 20, 2019)
[4] http://universaldependencies.org/ (accessed February 20, 2019)
[5] https://corpus.byu.edu/wiki/ (accessed February 20, 2019)

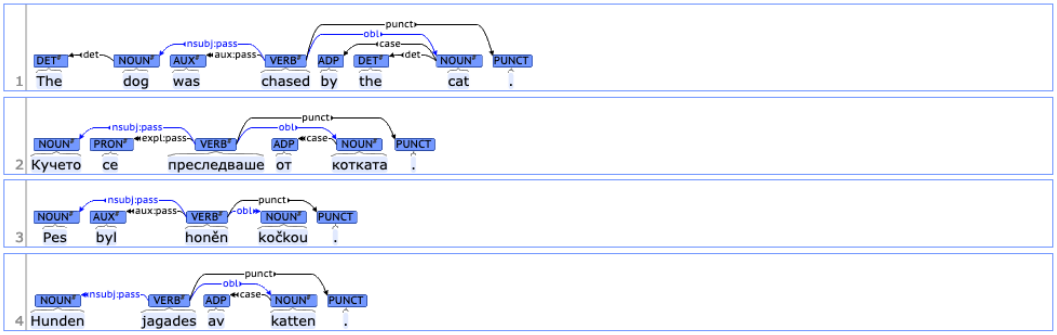

Figure 5 – Consistent annotation of similar constructions across languages (4).

With all these concepts in mind in the next chapters a proposed method will be detailed to solve a problem related to accessibility for DHH. The cited technologies will be the base for the design of the new method, which it will aggregate them aiming to compose a functional method.

3 Related Work

This chapter presents a survey, which covers some findings of the accessibility research using virtual and augmented reality systems, ranging years from 1996 to nowadays, and fields such as children, autism, motor rehabilitation, Parkinson disease, and inclusion for the impaired. Moreover, this chapter highlights the researches effort in DHH accessibility, which it's the main focus of this thesis.

Back in 1996 Virtual Reality fueled some good researches over the opportunities for child therapy, overall focusing on Autism Spectrum Disorders (ASDs) (39). Publications were collected up to 2005, when physical/motor rehabilitation raised up with new VR concerns. After that, later years came up with researches over rehabilitation for Traumatic Brain Injury (TBI), Parkinson Disease (PD) and the inclusion of impaired people. From that timeline division it was built this introductory rationale within scope groups, which are wherein as the following.

3.1 Children & Autism

This initial thread of studies tries to address the disability to learn everyday skills (40, 41) and develop cognitive abilities that ASDs professionals call 'executive functions' (42), amongst which are controls for interference and inhibition, integration across space and time, set shifting and maintenance, using planning, working memory, and social reciprocity (43).

A traditional debate exists between Behavioral and 'Theory of Mind' (ToM) methodologies (44), discussing their possible outcomes and presumed counter-parts in terms of affordability and generalization. The first methodology manages to offer better generalization results by applying an expensive 'prompt-fading' procedure (45), in which the children under-goes a deep process of reinforcement and repetition that requires a higher volume of applied resources (e.g. teachers, parents and schools). In the other side, ToM offers the child a repetition of tasks with lower complexity (44), managing to tech specific behaviors at a low cost without however succeeding to embody this knowledge down to a point in which the child starts to generalize (39), or in other words, find new matches of analogous situation in the real life. The main point is that a child with ASDs is capable of learning new rules, but unfortunately struggles to evolve it into social abilities.

In this context, some authors consider that VR technology is able to provide a safe environment for role-playing testing simulations, with slight and customized flexibility on the exercises to avoid the assimilation of static rules for specific situations and thus promote the generalization. VR is theorized to enable the results of a full implemented Behavioral methodology at low costs, with presumably better results (44, 46). However, studies are still

needed to scrutinize the possible levels of generalization and skill transfer allowed with the virtual environment.

3.2 Motor Rehabilitation

This accessibility thread reached results many times comparable or even better than those obtained via traditional rehabilitation therapy. Reports starting from 2003 mention a broad use of VR capabilities such as repetitive and hierarchical task administration, modulated task difficulty, and measurement of specific deficits and gains (47, 48). But still, prevalent advantage across studies seemed to be related with the greater enthusiasm that the technology evokes during the therapies.

A number of compared studies demonstrated reasonable success for specific VR rehabilitation programs (47, 49). Adults improved postural balance control, velocity and path deviation on virtual bicycle exercises, while achieving higher pain tolerance, cycling duration, overall distance and energy consumption (50). TBI studies of activity-based vs. VR-based exercises reported equal gains in the community balance and mobility scale, stating once more a greater enthusiasm within the VR group (51, 52, 53, 54); Exceptional levels of confidence and motivation were also achieved by users of VR-assisted orthopedic appliances and games (55, 56). Compared studies of occupational therapy also demonstrated better results with the VR technology, reporting greater levels of dynamic standing tolerance in geriatric patients (57).

3.3 Parkinson Disease

Parkinson Disease (PD) commonly affects patients motor and gait abilities, executive functions and attention capability with substantial impacts on the patient mobility, which is especially severe when associated with more complex, dual-task gait activity (DT) (58). Traditional PD therapy works to ease those symptoms and to promote better motor/cognitive functions. Some VR related reports however challenged this current paradigm.

In a research from 2010, twenty patients received progressive intensive Treadmill Training (TT) with virtual obstacles and revealed results that exceeded traditional therapy in several aspects. The authors managed to spot benefits of TT+VR on cognitive and motor symptoms of PD, dramatically improving DT costs by improving patient's attention management, enhancing gait speed and stride lengths with the development of extra motor and cognitive strategies to deal with virtual obstacle navigation. Gains for the proposed therapy were thoroughly evaluated along and thereafter in post study follow ups, confirming the transference, maintenance and even improvements out of the training sessions.

For all those findings, is it believed that VR can improve DT gait training by fostering motor adaptation and attention to the environment at superior levels, enhancing the ability to

learn new strategies and circumvent the impaired basal ganglia loops of PD within the daily life.

3.4 Inclusion of the Impaired

From 2012-2017, a new wave of researchers started studying VR as tool for the inclusion of impaired people. This accessibility thread targeted aspects of normal/functional life, proposing solutions that ranged communication (59, 60, 61), education (62, 63, 64) and entertainment (65, 66). Although none of the reviewed papers managed to define established standards so far, this sequel of reports is quickly learning from each other (61) while consumer industry for virtual and augmented reality (AR) becomes more and more accessible.

Authors are currently studying opportunities to apply VR into classrooms (62) and training centers (64), leveraging the capability of PWDS to successfully interact with both interlocutors (59, 60, 61) and machines (62, 64). As technology evolved to use real-time video streaming (64), mix accurate text-to-speech-to-text with noise reduction algorithms (59, 60, 61), apply automated visual saliency analysis (65) and advanced artificial intelligence (64) to communication, contemporary accessibility authors are continually proposing solutions and collecting participant results many times above the performance and satisfaction of 85% (59, 62, 65).

From Chinese deaf and hard of hearing (DHH) students using AR systems with virtual educators (62), DHH French communicating with normal people by using pseudo-phonetic live subtitles (61) all the way to Brazilians experimenting with VR/AR-based real-time theater captioning (67), and training centers for remote experimentation of electric powered wheelchairs (64). Accessibility keeps growing with resourceful directions and new research fields. After two decades of intense experimentation, accessibility in the VR stills in its infant and promising state.

3.4.1 Virtual Reality usage by impaired people

The inclusion of impaired individuals in virtual environments (VE) has been studied in the past years. In (68), mobility impaired users were studied aiming to check if by maximizing the presence in virtual environments users would be more effectively distracted from the pain and repetitiveness of rehabilitation, thereby increasing their motivation. The target of Guo et al. research was to understand how virtual environments affect users with mobility impairments(69). Specifically, the influence of full body avatars that have canes. This subject was further studied in (70), suggests that Persons with Mobility Impairments (PMIs) are easier to immerse in VEs than Persons without Mobility Impairments, which may further motivate the future use of VE technology for PMIs(71). Moreover, the impact of latency and avatars on perceived latency and gait parameters is investigated in (72). Samaraweera et al. present a study that quantifies the latency discrimination thresholds of a yet untested population - a specific subset of mobility

impaired participants where they suffered from Multiple Sclerosis - and compare the results to a control group of healthy participants(73).

In (66), an accessibility software prototype based on Samsung Gear VR Framework[1], which provides a framework to be used by developers, and has the purpose of adapting Zoom, Inverted Colors, Auto-reading (Screen Reader) and Caption features in a VR environment. The authors figure out users really enjoyed the application and suggested that other disabilities could benefit from it. This is according to defined goals within the study, once the target is to develop an application with tools that fulfill the needs of as many Persons With Disabilities (PWDS) as possible. If the application does not have any tools to help a specific disability, now it is possible to implement or adjust the existent ones. These recent researches serve as clear evidence of concern in bringing accessibility to virtual environments.

3.4.2 Virtual Reality as an assistive tool

These related works explored the VR technology to train, increase, or improve the functional capabilities of individuals with disabilities. Rodriguez presented a wheelchair simulator designed to allow children with multiple disabilities to familiarize themselves with the wheelchair(74). n this work(64), it is sustained the feasibility of a training environment for wheelchair users through a long-distance teleoperation that can be performed worldwide. With this system, it is possible to improve the quality of wheelchair training, creating user's immersion through a Head Mounted Display, and also allowing people with different level of disability to test and choose the alternative that most fits their capabilities. The objective of Cantu et al. research is to effectively design a cane interface for assistive and rehabilitative interactions in games(75). In (76), an assistive training tool for rehabilitation of dysphonic patients was designed and developed according to practical clinical needs. Furthermore, VR technology may be used as an initial step in the treatment of driving phobia, as long as it may facilitate the in vivo exposure, thus reducing risks and high costs of such exposure(77). Therefore, it might be able to explore the efficiency of using VR technology to help people with disabilities.

3.5 Virtual Reality solutions for Deaf or Hard of Hearing people

The use of Virtual and Augmented Reality to improve communication with DHH people in real-time is a chance for research with strong social impacts as it enables the social inclusion of impaired people in theater entertainment, conferences and all sorts of live presentations. This section highlights related works that followed this important thread.

[1] Accessibility is inside the Gear VR Framework (GearVRf) project, an open source collaboration based on the GearVRf open-source rendering library for application development on VR-supported Android devices, for more information: http://www.gearvrf.org/ and https://github.com/gearvrf/GearVRf-Demos/tree/master/gvr-accessibility

Mirzaei et. al. (59) solution improves live communication between deaf and ordinary people by turning ordinary people's speech into text in a device used by the deaf communicator. In this situation, the deaf also can write texts to be turned back into speech, so the ordinary people can understand. The solution is composed by a device and a software in which narrator (ordinary person) speech is captured by ASR or AVSR (Audio-Visual Speech Recognition) and turned into text, the Joiner Algorithm uses the text generated by ASR or AVSR and creates an AR environment with the image of narrator and text boxes of his speech. TTS engines are used to convert texts written by deaf people into speech for the narrator, making possible a two-way conversation. The results pointed that the average processing time for word recognition and AR displaying is less than 3s using ASR mode, and less than 10s for AVSR mode. To evaluate the solution, they conducted a survey with 100 deaf people and 100 ordinary people to measure the interest rate of using technological communication methods between them and 90% of participants agreed that the system is really useful, but there's still opportunity for improvements with AVSR mode, which is more accurate in noisy environments.

Berke (60) believes that providing word and its confidence score in a subtitle using AR environment, in order to give more information about the narrators speech, will improve deaf people understanding of a conversation. The solution proposal consists in a captioning which words generated by speech to text are displayed with its score of confidence in the subtitle and different colors are given for more confident and less confident words. These scores are calculated based on how sure speech to text algorithm are about the match of voice captured and the acoustic model of a word. The author also wants to study a way to present these information without confuse or make more difficult for the deaf to read the subtitle and pay attention on the narrator.

Piquard-Kipffer et. al. (61) made a study to evaluate the best way to present the text generated by speech to text algorithm in French language. The study covered 3 display modes: Orthographic, where recognized words are written into orthographical form; International Phonetic Alphabet (IPA), which writes all the recognized words and syllables in phonetic form by using the International Phonetic Alphabet, and lastly a Pseudo-phonetic where recognized words and syllables are written into a pseudo-phonetic alphabet. Some problems that challenge automatic subtitle systems such as noises captured by the device's unsophisticated microphones, implied in an incorrect word generation by ASR as reported in (59) thus flawing message understanding in deaf people's side. To minimize this negative, they included additional information about converted text within the subtitle for all display modes - like a confidence value for each word as proposed in (60, 61). Experiments with 10 deaf persons found best reviews when using a confidence score to format correct words in bold while presenting the incorrectly recognized ones in pseudo-phonetic mode; and suggested that preceding training phase for the experiment would be necessary to make participants more familiar with pseudo-phonetic reading. All participants manifested interest for such a system and thought that it could be helpful.

Hong et. al. (78) propose a scheme to improve the experience of DHH people with video captions, called Dynamic Captioning. It involves facial recognition, visual saliency analysis, text-speech alignment and other techniques. First, a script-face matching is done to identify which people the subtitles belong to in the scenes, this is based on face recognition, then a non-intrusive area is chosen in the video so that the caption can be positioned to avoid occlusion of important parts of the video and compromise the understanding of its content, the display of the caption emphasizing word for word is done through script-speech alignment and finally a voice volume estimation is done to display the magnitude indicator of the character's voice in the video . In order to validate the solution, the authors invited 60 hearing impaired people to an experiment that consists of watching 20 videos where some metrics such as comprehension and impression about the videos would be evaluated, in this experiment 3 captioning paradigms were tested: No Captioning, Static Captioning and Dynamic Captioning. The results showed that the No Captioning paradigm presented a poor experience for users, Static Captioning contributed to user distraction and 93.3% of users preferred Dynamic Captioning.

Beadles et. al. (79) patent propose an apparatus for providing closed captioning at a performance comprise means for encoding a signal representing the written equivalent of spoken dialogue. The signal is synchronized with spoken dialog and transmitted to wearable glasses of a person watching the performance. The glasses include receiving and decoding circuits and means for projecting a display image into the field of view of the person watching the performance representing at least one line of captioning. The field of view of the displayed image is equivalent to the field of view of the performance. A related solution for providing closed captioning further includes the step of accommodating for different interpupillary distances of the person wearing the glasses.

Luo et. al. (62) designed and implemented a Mixed Reality application which simulates in-class assistive learning and tested at China's largest DHH education institute. The experiments consisted in let these DHH children study a subject that is not in their regular curriculum and verify if the solution can improve the learning process. The solution has two main components, one component is the assisting console controlled by a hearing student, the other component is the virtual character displaying viewport which fulfills assistance. Both components use a dual-screen setup, within each component, one of the screens displays lecture video and the other screen displays mixed reality user interaction or rendering content. Videos on the screens of both components are synchronized in time. The hearing impaired student side of the system has a virtual character shown on the user content screen which can take predefined actions, while the hearing student side of the system has a control UI shown on the user content screen to manipulate virtual character at the other end to perform such actions. Results showed that the experience of being assisted by a virtual character were mostly very positive. Students rated this approach as novel, interesting and fun. 86,7% of them felt that with such help, it was easier to catch the pace of the lecture, understand the importance of knowledge points, and keep focused across the entire learning session.

Kercher and Rowe (63) propose a design, prototyping and usability testing of an AR head-mounted display system designed to improve the learning experience for the deaf, avoiding the attention split problem common among DHH people in learning process. The solution is focused in child's experience in a planetarium show. Their solution consists in three parts: Filming of the interpreter in front of a green screen, use a server to communicate the interpreter video to the headset and user interface for testing headset projection manipulation and optimization, then the interpreter will be always in the field of view of DHH spectator as it can also look freely to all directions and enjoy the show. The authors expect in the end of 3 years of research to not only help young children to have better experience in planetarium show but contribute in major changes in the experiences of the deaf in a variety of environments including planetariums, classrooms, museums, sporting events, live theaters and cinemas.

The Table 1 lists all cited researches for DHH accessibility powered by VR. Then, the table indicates the addressed topic of each research, as well as, this thesis proposal. The topics are DHH communication, Automatic Speech Recognition (ASR), Text Subtitling, Sign Language (SL) Subtitling, Speech Correction (SC), and Virtual and Augmented Reality (VR/AR) technology.

Researchers	DHH Communication	ASR	Text Subtitling	SL Subtitling	SC	VR/AR
Mirzaei at. al.	✓	✓	✓	✗	✗	✓
Berke	✓	✓	✓	✗	✗	✓
Piquard-Kipffer at. al.	✓	✓	✓	✗	✗	✗
Hong at. al.	✓	✓	✓	✗	✗	✗
Beadles at. al.	✓	✓	✗	✓	✗	✓
Luo at. al.	✗	✗	✗	✓	✗	✗
Kercher and Rowe	✗	✗	✗	✓	✗	✓
Thesis Proposal	✓	✓	✓	✓	✓	✓

Table 1 – Addressed topic of each cited research for DHH using VR.

3.6 Sing Language in VR Environment

There are a couple of works discussing about methods and challenges to create a signing avatar, as (80). Even as, (81), which describe the development of a new method of sign language subtitling using motion pictures. These works were studied to check how to add a signing windows to help DHH people to understanding the play in live theaters.

In AR/VR context, (82) created integrated multi-agent system involving a robot and virtual human designed to augment language exposure for 6-12 month old infants. (83) introduces the ImAc project, which explores how accessibility services (subtitling, audio description and

sign language) can be efficiently integrated with immersive media, such as omnidirectional and Virtual Reality (VR) contents, while keeping compatibility with current standards and technologies. These works give us a base to add a signing window in a augmented reality.

In (84), the authors explored the potential of AR as a novel way to allow users to view a sign language interpreter through an AR device while watching TV. In this cited work, the authors figure out by conducted experiments full-body format and the half-body format seemed the best two designs. All designs were placed on the right hand side of the TV frame. The authors cited the importance to maintain a 'connection' between the interpreter and the content, and in-turn the 'continuity' between the user, the interpreter and the content. Other important conclusion of this work is "The quality of the signing is more important than the presentation of the interpreter". This was a key when our proposed solution was designed.

These works (85, 86) describe the development and evaluation of a solution, which acts as a information support to deaf and hard of hearing people who are viewing sports programs using a sign language support system. The system automatically generates Japanese Sign Language (JSL) Computer Graphics (CG) animation and subtitles from prepared templates of JSL phrases corresponding to fixed format data. The authors concluded that the automatically generated JSL CG is practical enough for understanding the information.

Libras is the Brazilian Sign Language used by the Brazilian deaf community. Libras as any other sign language is perceived visually and is produced by gestures composed of movements of the hand, arms and body, combined with facial expressions. (87) describes a approach to build a comprehensive BP-Libras parallel corpus. The approach combines a methodology based on the translation of school textbooks with a thorough description of sign gestures and facial expressions based on motion captured data.

3.7 Implemented Systems to Enable Deaf or Hard of Hearing Accessibility

Even with all current technology, progressive innovators have failed to address some of the real problems that persist when it comes to accessibility. The problem to be solved by the proposed method described in this thesis is to bring accessibility in live theaters to Deaf or Hard of Hearing people.

In Butler (88), the author highlights that DHH viewers include advocacy for captions and caption formatting preferences; the need for direct access to real-time videos, online videos, and other media; how captions influence and benefit DHH and hearing viewers; and captions' importance in public, educational, and other social/cultural spaces. The author concludes that DHH viewers' perspectives can help educators and advocates strengthen access to captions in education and society. The proposed solution is focused on attending a part of this demand when

it aims to enable DHH people to access live theaters, a important and consolidate social/cultural space.

In the next sections are detailed the method from the sketch to a basic idealization. In the beginning of method concept, similar systems were researched to learn with these solutions. Moreover, a study with targeted audience was performed to collect proper feedback and requests from DHH community aiming to design a solution that attends the system users.

3.7.1 Movie Theater Solutions to Deaf or Hard of Hearing Accessibility

There is a lack of researches regarding DHH accessibility in live theaters, expect from solution described in section 3.7.2. Then, aiming to have a solid direction for this study, the movie theater environment was researched. There are a couple of attempts of bringing accessibility through VR technology for DHH people to movie theaters and other environments, as pointed out in chapter 3. In the next sections some of current effective systems are listed.

3.7.1.1 Rear Window Captioning System

The Rear Window Captioning (RWC) system is a technology that makes it possible for exhibitors to provide closed captioning for DHH moviegoers without displaying them to the entire audience. RWC is also significant because it doesn't require special OC prints or separate screenings, since the captions are not on the film itself. The Figure 6 shows the basic concept of the system.

The Motion Picture Access effort (MoPix) is an initiative of the National Center for Accessible Media (NCAM), a division of the WGBH Educational Foundation. This effort was launched in 1992 to research and develop ways of making movies in theaters accessible to deaf or hard of hearing people through the RWC system.

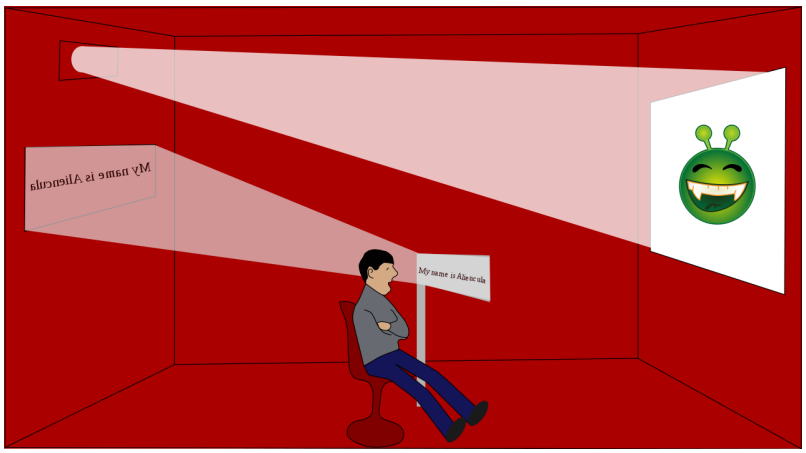

Figure 6 – Rear Window Captioning (5).

The proposed solution in this thesis needs to be different from Rear Window system due many theaters didn't have projection equipment.

3.7.1.2 CaptiView Closed Caption Viewing System

Doremi Cinema introduces the new CaptiView Closed Caption Viewing System for the DHH movie audiences. This system transmits and receives AES-128 encrypted closed captions on a wireless band frequency. With an 80 meter signal range, CaptiView can be used from ANY seat in the house (unlike existing "mirror image" systems that limit seat selection).

The CaptiView system consists of a small, OLED display on a bendable support arm that fits into the theater seat cup holder. The easy-to-read screen is equipped with a rechargeable Lithium Ion battery that lasts up to 16 hours per charge. The high contrast display comes with a privacy visor so it can be positioned directly in front the movie patron with minimal impact or distraction to neighboring patrons. The system is illustrated in Figure 7.

There are complains from DHH community that this kind of solution make it difficult to focus on both the film, which is in the background, and the screen, which is in the foreground. The constant shift in focus can be exhausting, and can also cause the viewer to miss a great deal of the action in the movie. Then, the proposed system will use VR technology to display the subtitling aiming to avoid this issue.

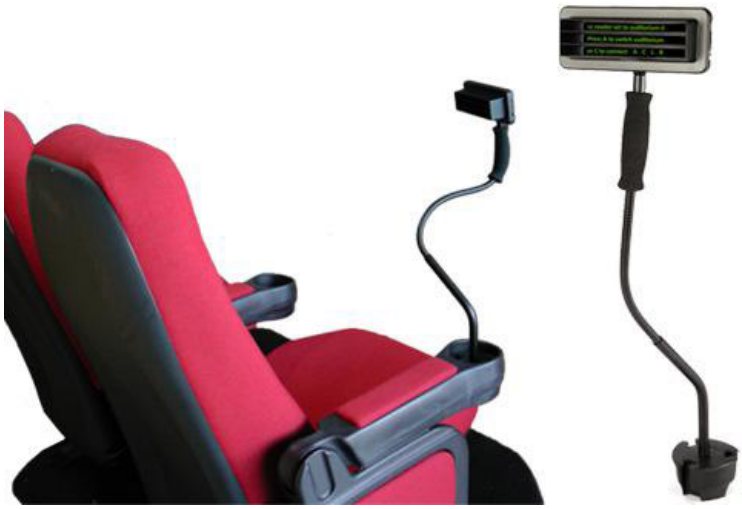

Figure 7 – CaptiView Closed Caption Viewing System for the DHH movie audiences (6).

3.7.1.3 USL Closed Captioning System

The USL Closed Captioning System (CCS) is designed to enhance the deaf or hard of hearing cinema patron's movie-going experience. A single infrared emitter broadcasts closed caption text and two channels of audio into an auditorium. Channel one is for hearing impaired (HI) and Channel two is for visual impaired narrative (VI-N). The use of IR instead of radio frequency transmission eliminates interference between adjacent auditoriums.

Two types of private display units are available: The "Seat Mount" display that clips to the arm rest and an "Eyewear/glasses" display. Each unit contains custom optics which display

the caption as a virtual image far enough from the viewer to avoid the need to refocus between the caption and the movie screen. The system is showed in Figure 8.

Figure 8 – USL Closed Captioning System (6).

3.7.1.4 Remarks

The Table 2 lists all cited systems for DHH accessibility in theaters. Then, the table indicates the addressed topic of each research, as well as, this thesis proposal. The topics are Automatic Speech Recognition (ASR), Text Subtitling, Sign Language (SL) Subtitling, Speech Correction (SC), and Virtual and Augmented Reality (VR/AR) technology.

System	ASR	Text Subtitling	SL Subtitling	SC	VR/AR
Rear Windows Captioning	✗	✓	✗	✗	✗
CaptiView Closed Caption Viewing	✗	✓	✗	✗	✗
USL Closed Captioning	✗	✓	✓	✗	✓
Thesis Proposal	✓	✓	✓	✓	✓

Table 2 – Addressed topic of each cited system for enable DHH people in theaters.

The proposed solution in this thesis is inspired in "Eyewear/glasses" display unit. The solution intends to improve the cited solution exchanging the USL property glasses display for the VR technology, which enables sign language window. Although, live theaters are a more complex scenario due actors timing and improvisation. All intrinsic characteristic of live theater must be taken in consideration in the proposed system, like play improvisation, which can be classified as: *a)* Substitution of some words of speech, *b)* Removal of entire speeches, *c)* Addition of speeches.

3.7.2 Open Access Smart Capture

A very similar system is been developed by National Theater (NT), located in London, UK, and Accenture company. The solution, called Open Access Smart Capture, uses smart glasses to enable a text subtitling to DDH people. Smart caption glasses are a revolutionary new way for people with hearing loss to enjoy performances at the National Theatre (93). By projecting captions in real time onto the glasses' lenses, DHH people can see the dialogue in front of their eyes. When wearing the glasses, users will see a transcript of the dialogue and descriptions of the sound from a performance displayed on the lenses of the glasses. This means that they can sit in any seat in the house, for any performance and still enjoy the show. Their only option before was attending the two or three performances in any production run where caption screens sat next to the stage. The smart caption glasses give anyone who is DHH the freedom to experience performances how and when they want to.

The glasses display a synchronized transcript of dialogue and sound from the production being viewed, directly onto the lenses of the glasses, giving viewers the freedom to experience performances how and when they want to. Following the launch of the glasses at the National, in 2019 the NT will partner with Leeds Playhouse as a next step toward helping to make this technology available in theaters across the entire U.K (94). The Figure 9 shows a picture of audience using closed-caption glasses in National Theater's

This solution seems very similar in proposed system in this thesis, but Open Access Smart Capture solution didn't cover sign language feature. Moreover, it couldn't be found any scientific publication about NT solution and its applied methods and implementation. Despite the similarities, the solutions are fully independent of each other form you concept and implementation.

3.7.3 Issues to be Solved

There are some studies related to subtitling in VR environment, like Roche (89), which states that there are three main issues which have to be considered:the position of the subtitles, the speaker identification and the influence for the VR experience. These issues were take in consideration during the system implementation.

The subtitles beginning with delay have most influence in the viewers' quality of experience as figured out by Guimarães (90), when the author states that subtitles synchronization (or the lack thereof) plays a key role, positively or negatively, in the perception of quality that viewers have about the content. Based on this finding, we focused on reduce the caption generation using sentence correction. In cited proposed method, once obtained the ASR output for an actor speech, a module for Speech Correction is used to identify in the play script a sentence that corresponds with what the Actor said, considering also the instant in the play in which the Actor expressed such speech. The result of this features addition is detailed in section 6.4.

Figure 9 – National Theatre's Innovative Closed-Caption Glasses (Photo: James Bellorini).

In Kafle (91), is cited that researchers who evaluate ASR performance often focus on improving the Word Error Rate (WER) metric, but WER has been found to have little correlation with human-subject performance on many applications. The author proposes a new captioning-focused evaluation metric that better predicts the impact of ASR recognition errors on the usability of automatically generated captions for people who are DHH. The accuracy of ASR technology has improved, but it is still imperfect in many settings. In this thesis will get this in consideration to proposes methods to improve the ASR performance for DHH people. In this thesis, a meticulous study is presented in section 5 to select the best ASP for this method implementation.

The latest Brazilian Census (92) states 5.1% of Brazilian population declares yourself as a DHH people, it's represents approximately 10 millions of people. Other important data from Brazilian census regarding DHH population is while 89.5% of the general population, aged 5 years and over, were literate, only 75.5% of the hearing impaired were and while 31.2% of the general population attended schools or daycare centers, only 12.3% of the hearing impaired did so. Based on this statistics, it's proved the importance of sign language addition in proposed solution due the fact of part of DHH population uses the sign language as main communication language. The study of sign language subtitling is described in chapter 4.2 and 7.2.

Aiming to evaluate the method, a quantitative and qualitative study were performed and it will be present in next chapters showing the results of the system regarding DHH understanding and satisfaction along entire play sessions.

3.8 Virtual Reality and Augmented Reality aptitude for accessibility

With the prior birds-eye overview of numerous accessibility fields - such as child autism, cognitive and motor rehabilitation, Parkinson disease, inclusion for interpersonal communication, training, education and entertainment - it became possible to summarize the overall VR and AR qualities in face of their impacts across multiple research outcomes. Four unique factors deserve mention in that sense: the motivational, the pragmatic, the adaptive, and lastly the distributional.

The motivational factor has been ubiquitously reported throughout all inspected areas of research, adding special value for the rehabilitation of the injured and diseased. This aspect greatly improved the enthusiasm and confidence of patients, but most importantly, it incited them to undergo with steadiness on the exercise repetitions, with superior adherence in the administered programs. This engagement level, particularly challenging for older patients, can come handy on many accessibility projects and studies.

Secondly and with nearly equal importance is the pragmatic factor. This one refers to results that VR and AR managed to deliver across distinct fields of research such as Parkinson disease, inclusion of several PWDS stereotypes, and at the rehabilitation as well. These results varied between comparable to preferable, sometimes achieving better results and succeeding to transfer/retain the acquired outcomes within patient's daily lives.

The adaptive factor had slightly lower importance so far but still holds good promises for the near future of studies, by when the complexity of implementations manages to increase. The possibility of integration with advanced technologies like real-time, automatic and AI-assisted feedback/evaluation algorithms can leverage results of accessibility research up to unprecedented levels, fostering the autonomy and performance of both users and professionals.

Lastly the distributional factor of VR/AR technologies might still take some time to influence accessibility results, probably when current research trials starts influencing industry grade products and services. Many authors across accessibility fields analyzed the importance of such characteristics of the technology, without however testing it thoroughly.

Relevant profiles of the technology will probably grow, evolve, and acquire new meanings over time. Still the snapshot of current achievements can be of an invaluable start for those wandering on VR/AR accessibility fields for the very first time.

Part II

The Proposed Method Overview

4 Subtitling Method to Deaf or Hard of Hearing People in Live Theaters

In this thesis is proposed a captioning system in Virtual Reality (VR) environment for DHH people aiming to improve their experience in live theaters (67) and include them socially, because it gives the opportunity for them to attend in live events.

The system works as follows: when an actor in a play speaks, the voice is captured by microphones, then this voice information is sent to an Automatic Speech Recognition (ASR) module to transform the voice into text. The ASR's outputted text is corrected by a module of speech correction that predates the play script for a syntactically or semantically similar sentence with that spoken by the actor. The output of the module of Speech Correction is then sent as a subtitle to all VR devices used by DHH spectators, allowing them to understand the play from the subtitles presented on a VR device they are using during performance. The component system working is showed in high level in Figure 10.

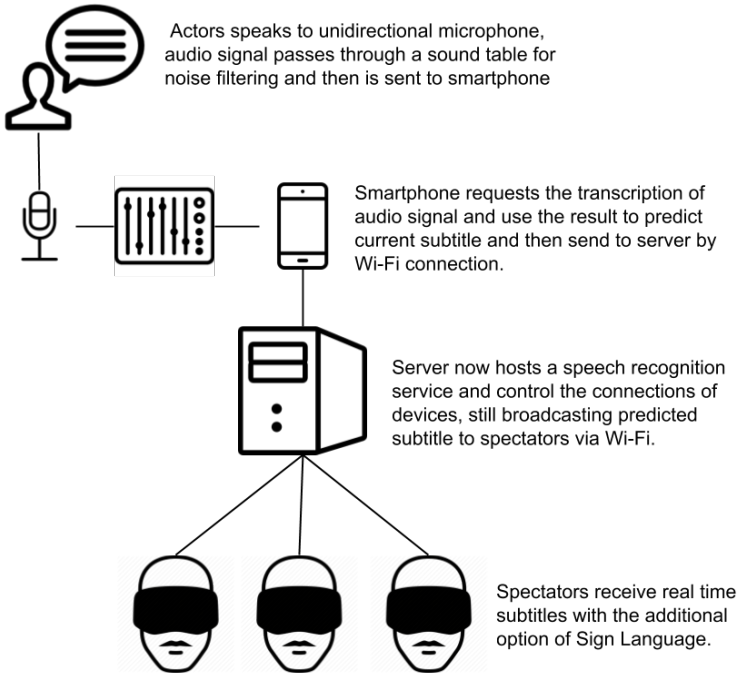

Figure 10 – Basic System Component Working.

Figure 11 presents the solution and it explains more specifically how system modules work.

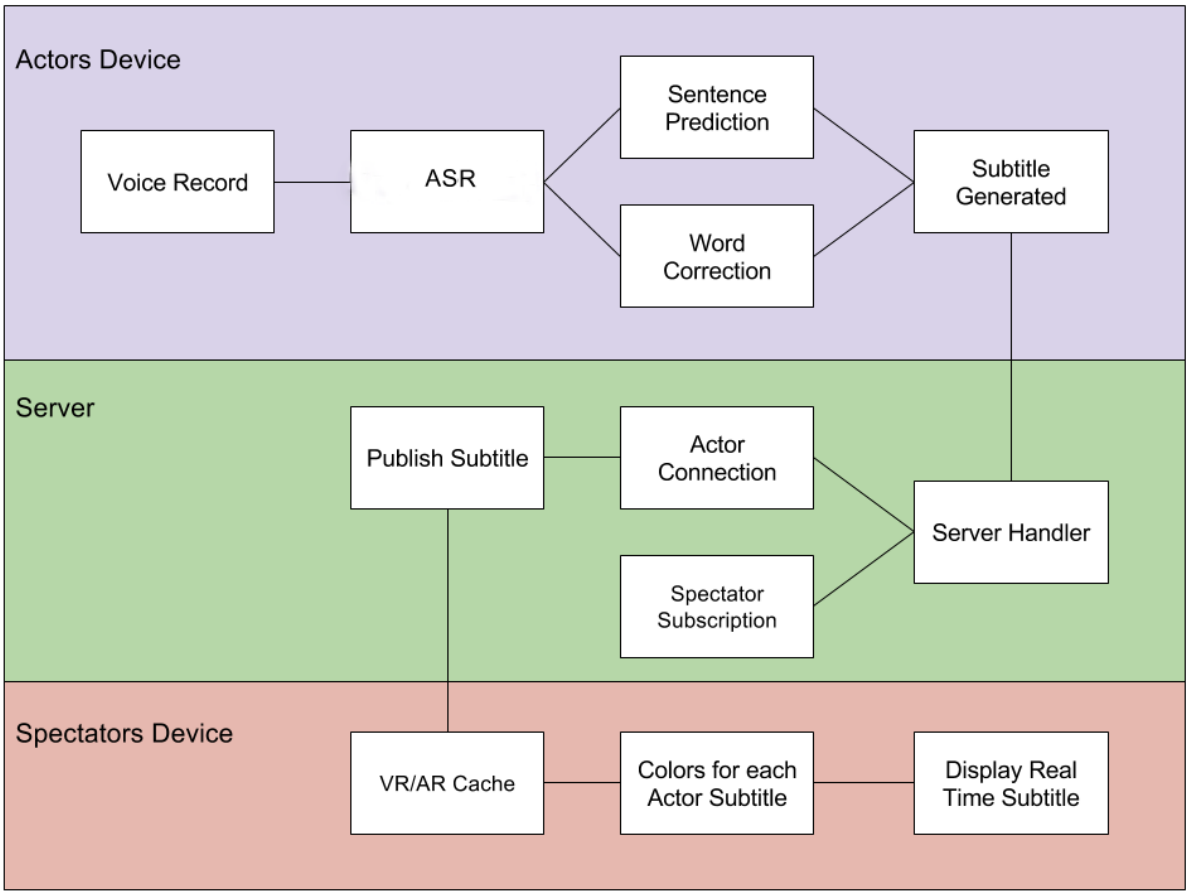

Figure 11 – Detailed Component System Solution.

Actors device: This module is composed by a) a Voice Record, where is used a unidirectional microphone; b) an ASR (Automatic Speech Recognition), a significant study about this component will be done to check the best features that fits with this proposed solution; c) Sentence Prediction and Word Correction aiming to improve the ASR results these components will be used to speech correction; d) Subtitle Generated will be use the play scrip and data from previous component to generate the properly subtitle in real time;

Server: Server module consists of a) Server Handler, which is responsible for device connection; b) Actor Connection, which controls the actor connection that can be n actors acting; c) Spectator Subscription, component where spectator requests to receive the subtitle; d) Publish Subtitle, where subtitle is sent to each spectator, which also can be n spectators.

Spectator Device: This module consists of a) VR/AR Cache, it's responsible for subtitle receiving; b) Colors for each Actor Subtitle, inside this component the subtitle is assigned a color, which represents a actor aiming to make easy the speech identification; c) Display Real Time Subtitle, the subtitle or sign windows is displayed in VR device.

Aiming to evaluate the method, a quantitative and qualitative study were performed and it

will be present in next chapters showing the results of the system regarding DHH understanding and satisfaction along entire play sessions.

In this chapter is described the proposed method in this thesis in details. Aiming to better expose the system, the chapter has two major sections, which are text and sign language subtitling. Then, the challenges of each subtitling mode is related, as well as, the implementation of each component in each mode.

Two approaches of the computer technique related to speech correction component are presented in this chapter, which are syntactic similarity and semantic similarity.

4.1 Text Subtitling Method to Deaf or Hard of Hearing People in Live Theaters

This section presents in details the proposed method, which uses an ASR and speech correction to retrieve the correct subtitle of live play scenes using text from play script. Figure 10 hows the concept model for system's architecture.

4.1.1 Actor Module

In Actors module, a ASR for Portuguese language converts recorded voice into text, using the first few words to retrieve the correct play speech using a sentence prediction algorithm before communicating it to server. The sentence prediction workflow is presented in Figure 12. When the device application starts, play script is submitted to Ngram algorithm which counts and calculates probabilities for all sets of N sentences, building a data table that serves as language model which can be queried with first N words converted from speech-to-text as key, and returning the most probable sentence.

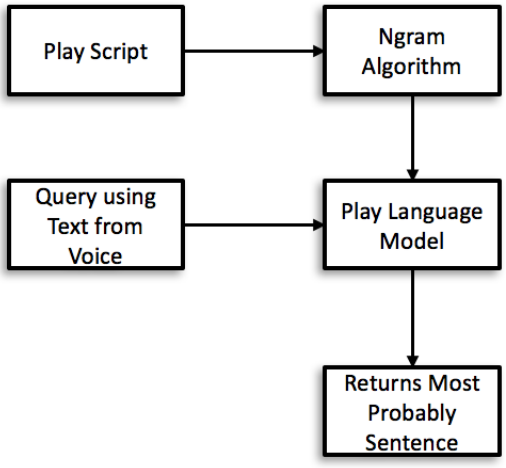

Figure 12 – System's sentence prediction.

Sentence prediction will ensure the behavior of retrieving subtitles in real-time while scene performance is occurring. If a search for speech play retrieves no match, system realizes that there is an improvisation occurring, and if needed it passes raw converted words to a word correction algorithm, which then communicates with the server. As demonstrated by Figure 13, word correction algorithm verifies if the words are present in Portuguese dictionary. If they are it skips correction, otherwise it searches for the most similar word in dictionary based on edit distance, and then overwrites those incorrect with similar words. Incorrect words are saved to be used as data for word correction retraining, as well as words captured from improvisation, which can be used in future sentence prediction retrain. Subtitles sent to server contain actor's character identification, line of speech identification and the subtitle itself. Server uses these line values to broadcast subtitles in the correct order. Lines for improvisation messages receives an special value.

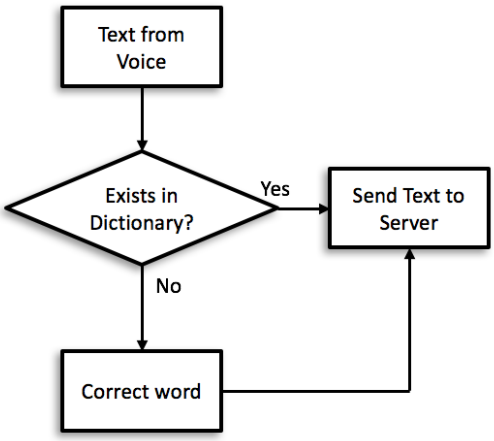

Figure 13 – System's word correction.

4.1.1.1 ASR Module

In the beginning of this solution implementation, the system consumed the ASR engine provided by the Google Cloud Speech-to-Text API[1], from now on referred as Google ASR. Although this API has a good quality of speech recognition, it depends of an Internet connection. Due to connection difficulty inside many theaters in Brazil, consuming from services on the Internet introduce several problems associated with delays and lost of data. To resolve this problem, we aim to replace the Google ASR with another that allows processing speech recognition in offline mode (i.e without need of internet connection). An ASR system was provided by the brazilian institute Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPqD). It is of interest for this work to evaluate such module in order to determine feasibility of its use as a replacement of the current ASR being employed.

[1] https://cloud.google.com/speech-to-text

4.1.2 Server Module

All performing actors in stage for the play carried a unidirectional microphone connected to a Samsung S8 device. Unidirectional microphones prevented undesired recording of noises from environment and voices from other actors. Each device had a copy of the play script to support the task of retrieving actor's speech as text. Actors voices were continuously recorded by their devices and no further interaction was required to operate the application.

All actors and spectator's device connections are managed by a single server, which broadcasts subtitles received from all actors to all spectators. The first few words converted into text are used to search for the exact play speech using speech correction before communicating it to server. If there is no match, the system understand it as an improvisation, and starts sending every word to server as soon as they are converted from actor's voice.

The server has a control panel screen with some basic functions designed to an operator follows the script play, identify the subtitling status transmission in each spectator device, identify errors in actor module, and send manually some texts from script play to the spectators. The settings screen is showed in Figure 14. A loaded script play is illustrated in Figure 15 and a play in execution is showed in Figure 16, green highlights indicates these speeches were sent to spectators devices.

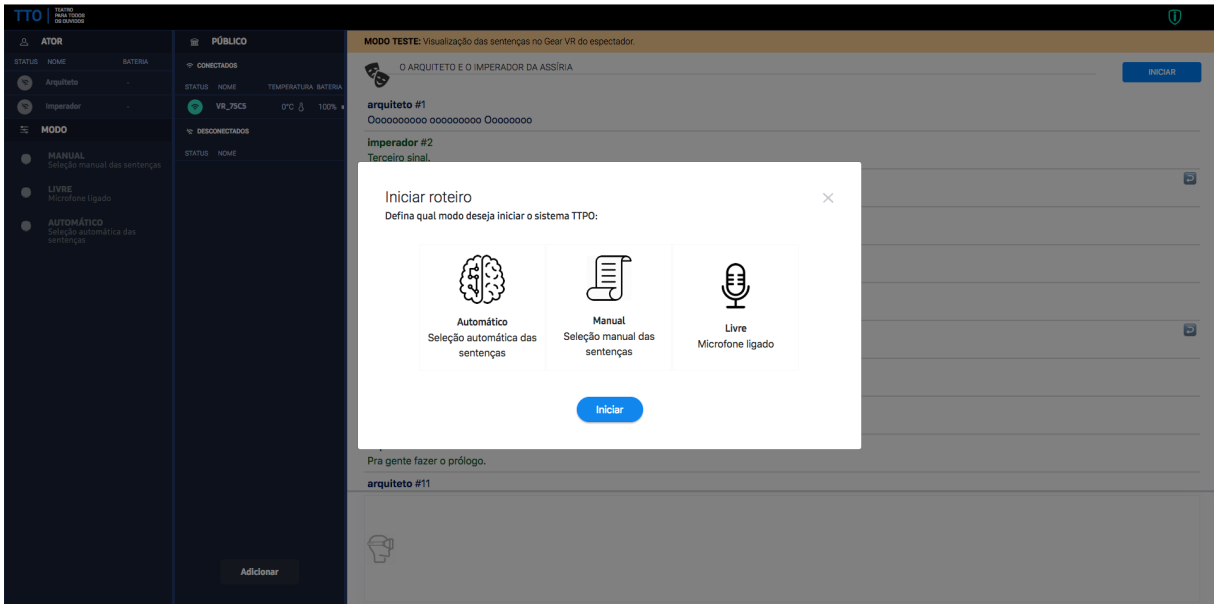

Figure 14 – Settings of server control panel screenshot.

4.1.3 Spectator Module

Server have handlers for actor devices connections, disconnections and subtitle communication. Each actor's device has its own connection which remains open until application closes. In the other side, spectator devices do not send messages to server, using only available

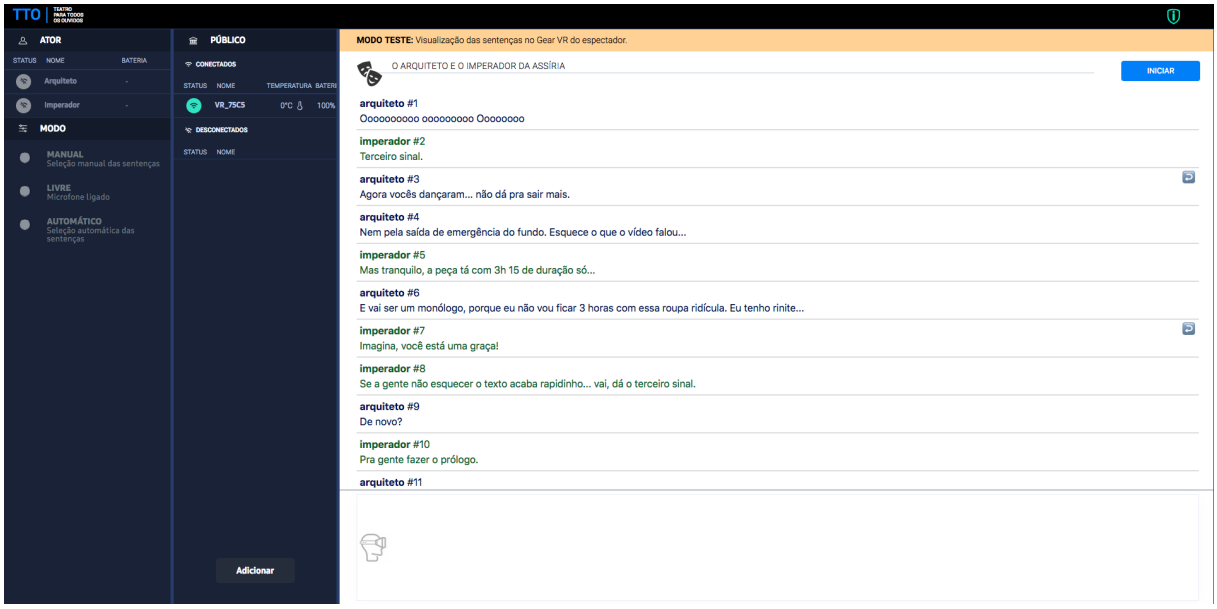

Figure 15 – System ready to follow a play in server control panel screenshot.

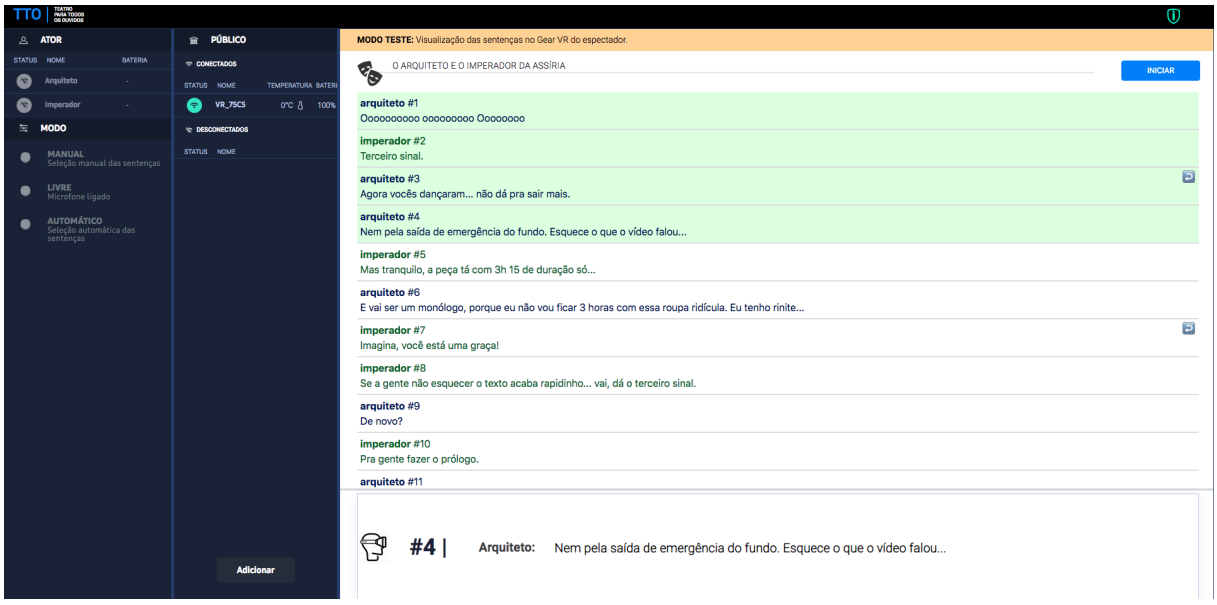

Figure 16 – Play running in server control panel screenshot.

subscription services. For each subscribed spectator, server broadcasts every received subtitle until unsubscription request happens. When subtitles are received by AR/VR application, they are added to a queue and wait to be displayed. The application renders each actor's subtitles with a distinct color that is consistent during all play. When presenting these, a calculation is made to verify whether text fits in the UI space, and if they don't, subtitle gets split and displayed by parts. Then a comfortable time estimation is given based in each subtitle's size, defining when it will be overwritten by the next subtitle. When there is no subtitles in queue, application simply listens for upcoming server subtitles until the end of the play. Figure 17 shows an demonstration of how subtitle is in VR environment.

Figure 17 – Subtitle in Gear VR.

4.2 Sign Language Subtitling Method to Deaf or Hard of Hearing People in Live theater

In this section, the study detailed in chapter 4.1 with DHH people on theater environment was continued aiming to improve topics which was bad rated on user tests based on collected feedbacks, like the performance of speech correction when an improvisation occurs. It is implemented Sign Language (SL) as an additional option of subtitle in order to turn the prototype more inclusive to DHH people. Then, two challenging tasks was solved, which were find an algorithm more resilient to improvisations and find the best way to display a window containing a sign language interpreter in VR environment without compromising the spectator view of the play and its understanding of the content.

During tests on theater using prototype of last study, It was observed that improvisations were more frequent than expected. Improvisations can be classified as following:

1. Substitution of some words of speech: Happens when actors forget the exact sentence to be said, according to play script.

2. Removal of entire speeches: Happens when stage director realizes actor is missing too often certain speech or when it is needed to shorten the time of play.

3. Addition of speeches: Happens when stage director wants to increase the time of play or to add extra feeling to current scene.

Since the speech correction is based on NGram algorithm, it is expected from the system to miss the speeches which matches with any of described types of improvisations, because NGram makes string comparison internally. For this reason the core of speech correction was changed to use word embeddings and edit distance. This combination is called semantic similarity. This approach was choosed because word embeddings represents the semantic of a single or more words as a vector in feature space, so this solution no longer have to worry about if words in a speech are misspelled or changed.

The system architecture is shown in Figure 10 in high level design. The next sections will describe each component of architecture. The components were redesigned to enhance the solution detailed in chapter 4.1.

4.2.1 Actor Module

In Actor Module, the actors voice passes through a sound table for noise filtering before being processed by Android application. After that, the application uses this audio signal to request its transcription from a local ASR service hosted in Server Module and then the transcription is submitted to speech correction.

On speech correction, the transcription and each of the play speeches in the window of speeches is tested using the new approach of word embeddings like shown in Figure 18.

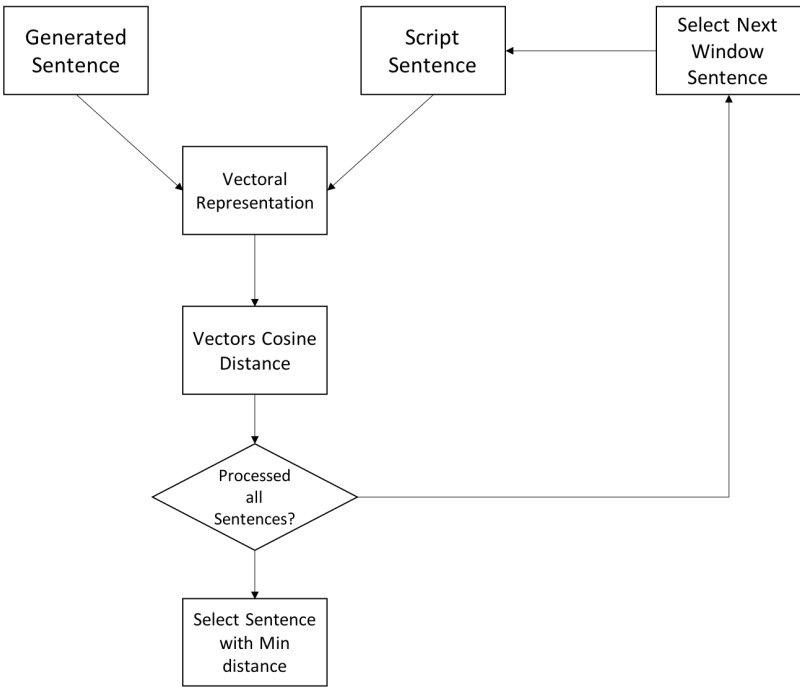

Figure 18 – Prediction of current speech using word embeddings approach.

Both transcribed text and current play speech of window are converted to vectoral space, then we calculate cosine distance between the vectors to measure the similarity between both sentences and then store the score, this calculation will be made to all play speeches in the

window one per loop. after calculating the scores, application will choose the lowest value, that means the closest vector and thus the most similar sentence.

Even when a type 1 improvisation comes to application, it is expected that SP still performs good because in type 1 improvisations the words change but the sense remains the same and since word embeddings compares the semantics of sentences and not the words, these improvisations will still be close to some play speech in vectoral space and thus the application can process these like other sentences. After creating the subtitle, it is sent to server via Wi-fi connection as usual.

4.2.2 Server Module

Server Module has a new service which is an local STT server for Portuguese language, so smartphones in Actor Module no longer need to Internet connection to perform the actors voice transcription, this speeds up subtitle generation. We also implemented a an Administration page which a technician can send manually subtitles in case of a smartphone in Actor Module fails and watch spectators devices battery, temperature and connectivity status, warning a theater staff that a spectator device needs to be changed. The message to be sent to Spectator Module has an additional field which is the timestamp (begin and end) of Sign Language video corresponding to the translation of current message, this new format will provide support to Sign Language display in Spectator Module.

4.2.3 Spectator Module

Spectator Module now has a menu which user has Sign Language (SL) as one of subtitles options. Each of play speeches is linked to a video of its translation in Sign Language and the videos are stored within Unity VR application. The application receives server message containing text of chosen play speech and timestamps to make possible the change between subtitle options when desired. When SL option is enabled and receives a message from server, application renders the video of translation which is a window containing a SL interpreter is shown like in Figure 19.

In order to make the interpreter to transmit a sensation of something included to VR environment, OpenCV framework was used to perform background subtraction on the sign language translation, removing the chroma key and leaving only the interpreter

Translation video is always displayed even when finishing to play all subtitles in queue, but instead of disappear, it will stay in an idle position, disappearing only if text subtitle is chosen. The Figure 20 illustrates the idle position. Moreover, it's showed the original translation video captured using a chroma key background.

To produce the sign language translation video, first an interpreter went to an exhibition of play "O Arquiteto e o Imperador da Assíria" to understand what is the play's mood and which

Figure 19 – VR view of Sign Language subtitle option enabled.

Figure 20 – Translation video in idle position.

signals to choose in order to convey the correct feelings of scenes, after that the translation video is recorded using a chroma key background and submitted to an invited group of DHH people to validate the clarity of translation. Then, the interpreter marks timestamps corresponding to each speech of play.

From UX side, the core challenge for proposing it was the lack of mentions related to accessibility guidelines for sign language applied on AR. From deaf people perspective, due to factors as price of equipment, people do not know properly how AR technology performs which increase the effort to comprehend their mindsets and expectations in a way to create an intuitive

user interface.

Hence it was opted starting the UX guidelines based on parameters for Web and TV added to a collaborative approach with a local deaf community. TV and Web parameters provided the initial recommendations of the sign language window setup: dimensions, contrast, and position. It was observed that most of solutions utilize Portuguese subtitles only (95). Unfortunately it attends partially the DHH needs because most of them have proficiency in LIBRAS (Brazilian Sign Language). They need support when exposed to writing language. The collaborative approach was used to investigate positive and negative aspects of the current solutions. The results showed that there is no consensus about preferences for dimensions, positioning or background color.

The initial studies resulted in two sizes of sign language window. The position of components considered the Eye comfort Zone recommended by Google DayDream (96) added too smartphone field of view. After adjustments a short video was applied in front of a static background which simulated a theater environment illustrated in Figure 21. The options were evaluated by 7 DHH people. All of them said the position on the right side was adequate. Young people preferred the small one and old people preferred the normal size. When asked about the best solution they mentioned that customization features tends to be the best way to attend properly all preferences.

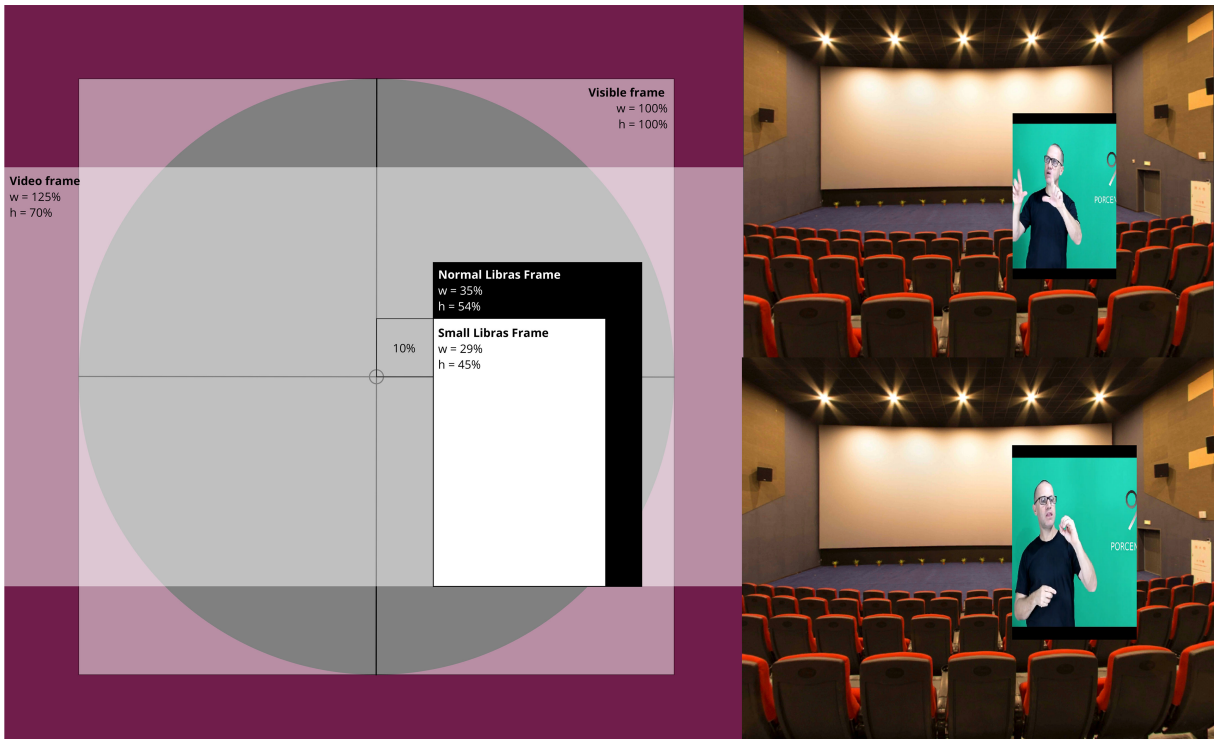

Figure 21 – First version of sign language window.

Despite the initial results indicate the right side as appropriate side to arrange the sign language component it was observed no consensus for some definitions like color background and size of window. Aiming to verify with end users the satisfactory parameters for those subjects

an interactive proof of concept was built. This model allowed users to choose their preferences for color background, opacity level, window size, and horizontal position. This is illustrated in Figure 22. Each category presented at least 3 options to be analyzed. 6 DHH people evaluated this model. At the end the following parameters were defined: for setup in Unity 3d: Width x Height: 1024x1024 (Gear VR, Oculus Render Eye Texture Size, Position: x, y, z (0, 0, 10), Canvas: Plane Distance: 10, Canvas Scale: Scale Factor: 1, Video (Libras) Renderer Transform: Position: x, y, z (64, -70, 0), Width x Height: 184x184; in a 2D space frame Size: 1024x1024, frame position: 0,0 (center), "Libras", Size: 184x184, Position: 64, -70.

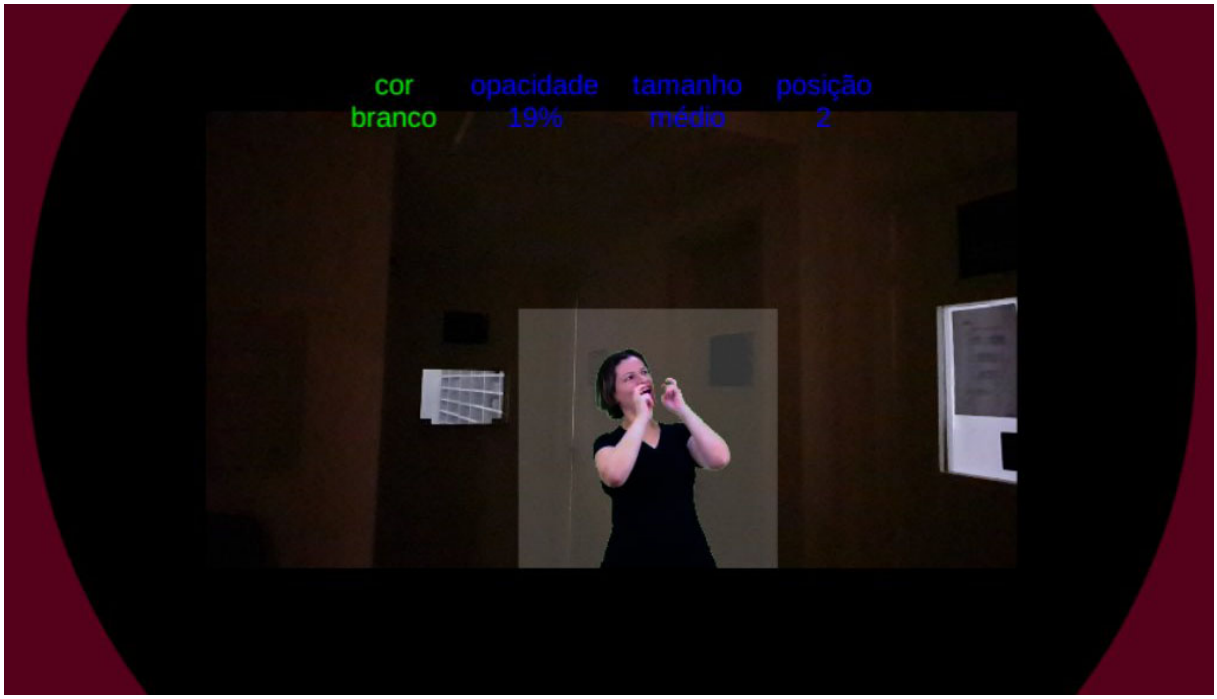

Figure 22 – Interactive proof of concept for sign language.

4.3 Considerations

With all method exposed in this chapter, the proposed method overview part can be ended. Then, the next part of this thesis can be started, where the method will be fully analyzed.

In the next part of this thesis, part III, Analysis, it's covers the analysis of the proposed method in several point of views. From ASR usage point of view is showed in chapter 5. From Speech Correction component perspective is detailed in chapter 6. From DHH users perspective is showed in chapter 7. The analysis will summarize the performed experiments, as well as, all results obtained from them.

Part III

Analysis

5 Experiments and Results related to Automatic Speech Recognition

This chapter presents a comparison between the Google ASR API against the CPqD ASR server solution based on four metrics, namely: a) Word Recognition Rate, b) Word Error Rate, c) Real Time Factor and d) Cross-entropy. It is also purpose of this comparison to promote information that permits identify the pros and cons of using offline mode provided by the CPqD ASR server. These two ASRs were used in this proposed method. This evaluation aims to decide the best ASR for this solution.

5.1 Experiment setup

Figure 23 presents the architecture employed to collect variables that describe the process of automatic speech recognition performed by the ASRs. During this process each audio file in the corpuses is submitted to the engine of speech recognition and is expected a response from which is possible to retrieve a recognized text string as well as a score of confidence that the engine assigns to the recognition.

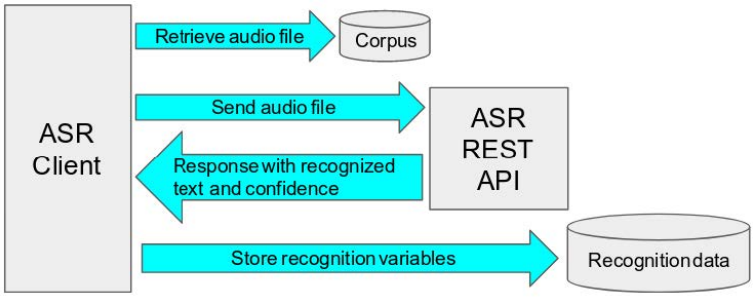

Figure 23 – ASR recognition data retrieving.

The process of audio submission for speech recognition as implemented in this evaluation and as expected to be implemented in the actual solution, follows the client-server architecture. For each ASR evaluated was implemented a client that attends to specifications of the communication interface between a client and the ASR. Such interface is implemented as an API that follows the HTTP protocol and provides a functionality that abstracts submission of an audio file to the ASR engine and retrieve the output of recognition (i.e recognized text and confidence of recognition).

The client application employed to consume from the Google ASR was implemented in the python language version 3 using the SDK publicly available provided by the authors of such ASR. To consume the API of CPqD was implemented a client application using python version

3 that performs HTTP requests following instructions and a sample client application provided by the authors[1].

If the audio file submission succeeds and a correct response is obtained, a set of variables are gathered from this process. In Table I are presented the variables gathered during each audio file submission, as well as its description. Once all audio files are submitted and the defined variables' values are gathered, such information is employed to calculate a set of metrics that permits evaluate performance and accuracy of the ASR. In coming sections are explained characteristics of the datasets employed as well as the metrics lifted for evaluation.

5.2 Datasets

Two datasets are employed to perform ASR evaluation, namely VoxForge and Laps-Benchmark. Such datasets contain audio and text files in which for each sentence in the text files exists an audio file of a person speaking such sentence. Following in this section are presented further details on the datasets.

5.2.1 The Laps-Benchmark dataset

The Laps-Benchmark dataset is recommended by the research group FalaBrasil[2] of Universidade Federal do Pará that posses several works in the field. Such dataset was obtained at the official Github space of the group[3]. It consists of audio files recorded by 35 persons distributed in 25 men and 10 women. Each person recorded 20 quotes in a separate audio file by quote, totalizing 700 quotes and approximately 54 minutes of audio. All recordings were performed using a computer with a regular microphone in a not controlled environment with the presence of ambient noise. The records in the original dataset are stored in Mono channeled, uncompressed 16-bit PCM audio files in format WAV with a sample rate of 22.500 Hz, however the audio files are down sampled to 16.000 Hz to attend requirements of the ASRs evaluated.

5.2.2 The VoxForge dataset

The audio files in the VoxForge dataset[4] are obtained from an Open Source initiative with the same name, created with the intention of collect several audio of people worldwide, speaking in its native language, to be employed on the training of ASR engines. The audio employed in this evaluation is a subset of the available audio in the official site of VoxForge and belongs to anonymous brazilian portuguese speaking persons that accessed the site and accepted the request of recording and audio file voluntarily. Each person recorded approximately 10 randomly

[1] https://speechweb.cpqd.com.br/asr/docs/latest/get started/index.html (accessed February 10, 2019)
[2] http://labvis.ufpa.br/falabrasil/ (accessed February 10, 2019)
[3] https://github.com/falabrasil/corpora (accessed February 10, 2019)
[4] http://www.repository.voxforge1.org/downloads/pt/Trunk/Audio/ (accessed February 10, 2019)

selected quotes in separate audio files. No information was obtained on the characteristics of the recording devices employed or about the environment in which the audio was recorded. Such records are stored in Mono channeled, uncompressed 16-bit PCM audio files in format WAV with a sample rate of 16.000 Hz.

5.3 Evaluation metrics

In order to evaluate the quality of the ASRs to be included in the module of speech recognition, a set of metrics are lifted from the parameters observed on the batch of submissions executed following the setup presented in section 5.1 . Such metrics were selected given its use in previous research works that assess evaluation of ASR systems including brazilian portuguese speech recognition (97, 98, 99).

5.3.1 Word Error Rate

In most ASR applications the figure of merit of an ASR system is the Word Error Rate (WER) (99). The WER is a proportion of how many errors exists on recognition against how many words exists on the reference, based on: a) substitutions, the words in same positions of hypothesis and reference are different; b) deletions, words of the reference does not appear in the hypothesis; c) insertions, different words than those in the reference are included in the hypothesis. In equation (5.1) is presented the formula employed to calculate the WER.

$$d = \frac{S+D+I}{N} \tag{5.1}$$

where: S = substitutions
D = deletions
I = insertions
N = size of reference

5.3.2 Word Recognition Rate

The Word Recognition Rate (WRR) can be intuitively thought as the accuracy of recognition. This metric is the proportion of correctly recognized words over the size of the reference. The formula employed to calculate the WRR is presented in equation (5.2).

$$WRR = \frac{C}{N} \tag{5.2}$$

where: C = correctly recognized words
N = size of reference

5.3.3 xRT

Another metric for evaluating an ASR system is the realtime factor (xRT). The xRT is obtained by dividing the time that the system spends to recognize a sentence by its time duration (99). A lower xRT indicates a faster recognition. In equation (5.3) is presented the formula employed to calculate this metric.

$$xRT = \frac{RT}{AD} \tag{5.3}$$

where: RT = recognition time
AD = audio duration

5.3.4 Cross-entropy

It is also interest of this study to evaluate ASRs' confidence on recognition. It is understood that prediction systems are more trustworthy if present a high certainty, i.e the probability the system gives of its prediction being correct is among 90%.

The certainty of an ASR is reflected from the probabilities it gives of a word or groups of words appear after another in a sentence, which is expressed in a structure called the language model. More accurate and with fewer uncertainty language models gets to represent in its internal structure a prediction with lower values of entropy. Such entropy is possible to estimate from the probabilities provided by the ASR on recognition.

$$Hp(T) = -\frac{1}{W_T} log_2 p(T) \tag{5.4}$$

where: T = recognition outputted text
$p(T)$ = probability of prediction
W_T = size of T

A metric that permits evaluate certainty of prediction is the cross-entropy. Intuitively the cross-entropy is the average entropy of recognition probability from each word in the outputted string. In equation (5.4) is presented the formula employed to calculate the cross-entropy.

5.4 Results

In this section are presented the results of experimentation following the setup in section 5.1. All plots provided in this section correspond to the metrics proposed in section 5.3. Every plot correspond to a metric measure obtained by both ASRs in a specific dataset. Intentionally

on each plot is expressed a metric evaluated for both ASRs on a given dataset in order to induce a comparison.

The graphs present how recognitions are distributed in the scores obtained. The bars' eight represent which percent of the results are located among the bounds defined by the bins in the horizontal axis. All bins in a single graph have the same size. The leftmost bin always start in the minimal score value encountered in all results and the rightmost bin always end with the maximum score value encountered, for any given metric in all datasets. The quantity of bins for each graph were selected so as the results are relevant for evaluation. Bins with zero instances are not plotted.

Figure 24 presents the distribution of sentences by Word Recognition Rate scores presented by both ASRs when applied on the VoxForge dataset. Each bar represents how many sentences (i.e percent of dataset) the ASR presented a WRR among the start and ending value of the bin such bar is located.

As observed the Google ASR recognized 10% more sentences of the VoxForge dataset than the CPqD ASR in the same dataset with a WRR above 0.8. If established a threshold of sentence understandability of 80%, approximately 1000 sentences outputted by the CPqD would have not been understood and 700 in the case of Google. Approximately 155 sentences was completely wrong recognized (i.e no word of the sentence was correctly recognized) by the CPqD ASR, 108 more than the Google ASR.

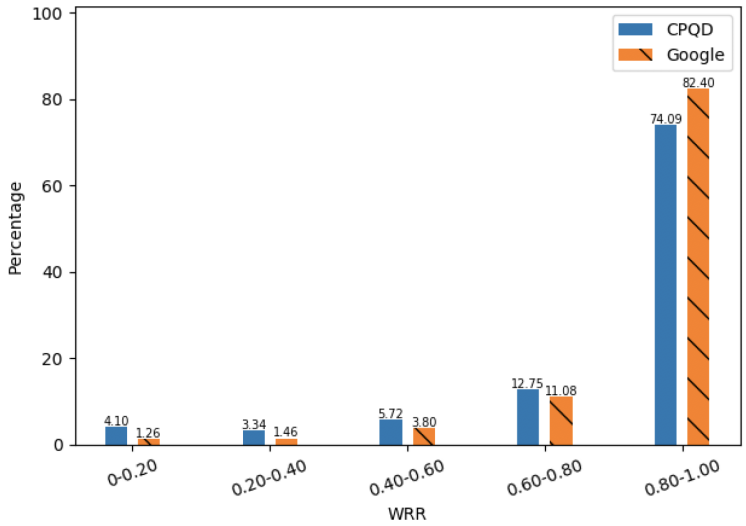

Figure 24 – Google vs CPqD Word Recognition Rate on VoxForge dataset.

Figure 25 presents the distribution of sentences by Word Recognition Rate scores presented by both ASRs when applied on the Laps-Benchmark dataset. Each bar represents how many sentences (i.e percent of dataset) the ASR presented a WRR among the start and ending value of the bin such bar is located. In the case of the Laps-Benchmark dataset while the accuracy of the Google ASR enhanced, the CPqD ASR stay equal or reduced if considered that the bins

have not the exact same size than in Figure 6. No ASR presented completely wrong recognized sentences, at least 10% of any sentence was recognized for the CPqD ASR and 28% for the Google ASR.

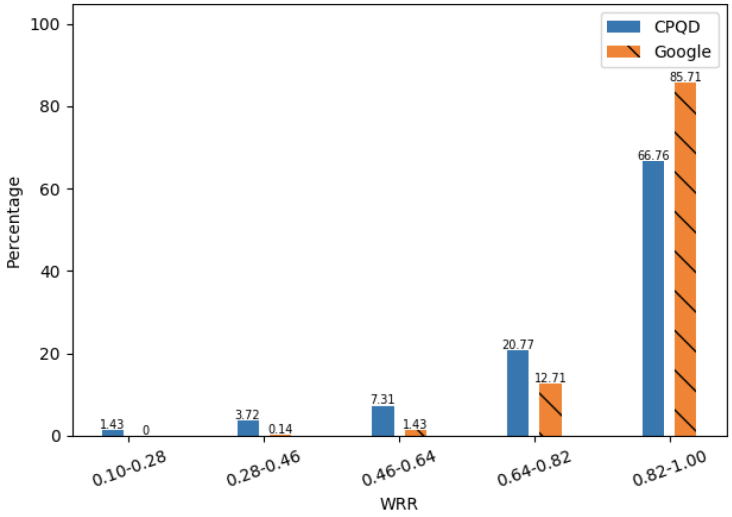

Figure 25 – Google vs CPqD Word Recognition Rate on Laps-Benchmark dataset.

Figure 26 presents the distribution of sentences by Word Error Rate scores presented by both ASRs when applied on the VoxForge dataset. Each bar represents how many sentences (i.e percent of dataset) the ASR presented a WER among the start and ending value of the bin such bar is located. For approximately 450 sentences the quantity of errors introduced in the recognition of sentences in the VoxForge dataset by the CPqD ASR (i.e quantity of substituted, inserted or deleted words) exceeded the size of the reference. Even though as seen in Figure 26 there are sentences that are entirely recognized wrong it does not mean that for all such 450 sentences part of the sentence was not correctly recognized, however subjectively such sentences can be considered as hard to understand or associate with another predefined text based on its similarity (as is the case of the solution proposed in this thesis).

Figure 27 presents the distribution of sentences by Word Error Rate scores presented by both ASRs when applied on the Laps- Benchmark dataset. Each bar represents how many sentences (i.e percent of dataset) the ASR presented a WER among the start and ending value of the bin such bar is located.

Less errors are introduced on recognition by both ASRs when applied on the Laps-Benchmark dataset. For all sentences the quantity of errors did not exceed the quantity of words in the sentence. The Google ASR did not introduced incorrect recognized words that exceed more than 63% of any sentence.

Figure 28 presents the distribution of sentences by real time factor xRD scores presented by both ASRs when applied on the VoxForge dataset. Each bar represents how many sentences (i.e percent of dataset) the ASR presented a xRD among the start and ending value of the bin

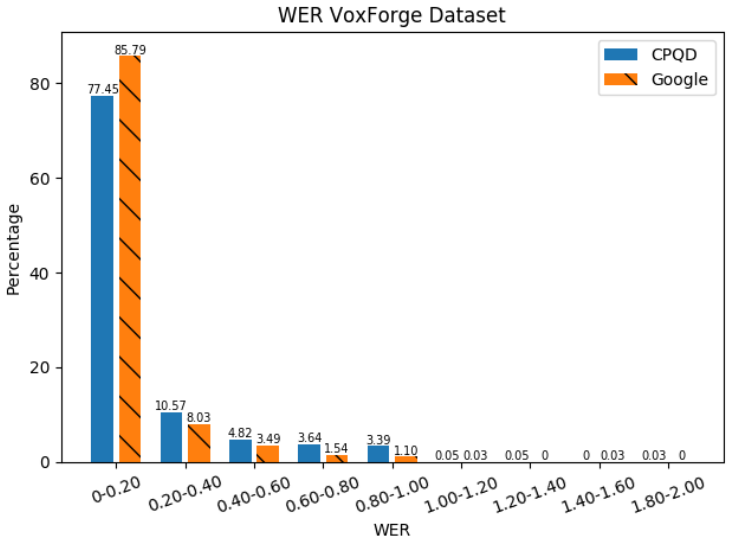

Figure 26 – Google vs CPqD Word Error Rate on VoxForge dataset.

.

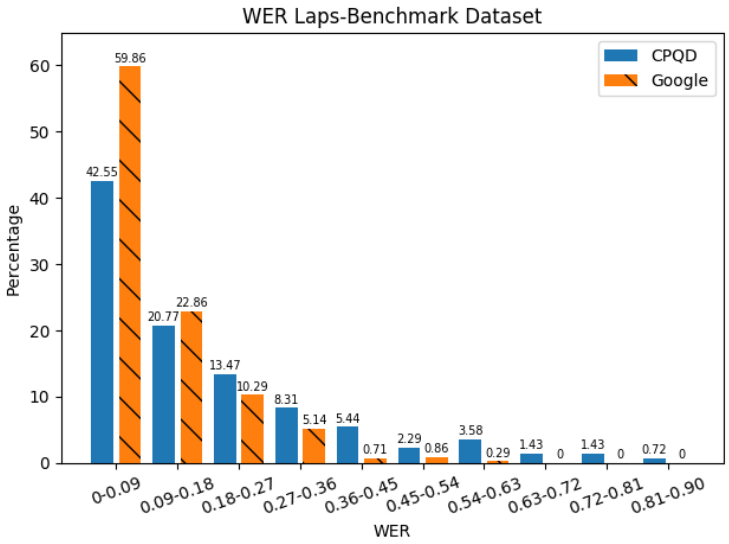

Figure 27 – Google vs CPqD Word Error Rate on Laps-Benchmark dataset.

.

such bar is located.

As observed in Figure 28 the CPqD ASR did not take more than the half of the duration of any audio file in the VoxForge dataset to output a recognition. If considered that the average duration of an audio file in that dataset is 3 seconds the CPqD ASR responded in an average of 1.5 seconds. In some cases the Google ASR took almost 3 times the duration of the audio file to output a recognition.

Figure 29 presents the distribution of sentences by real time factor xRD scores presented by both ASRs when applied on the Laps-Benchmark dataset. Each bar represents how many

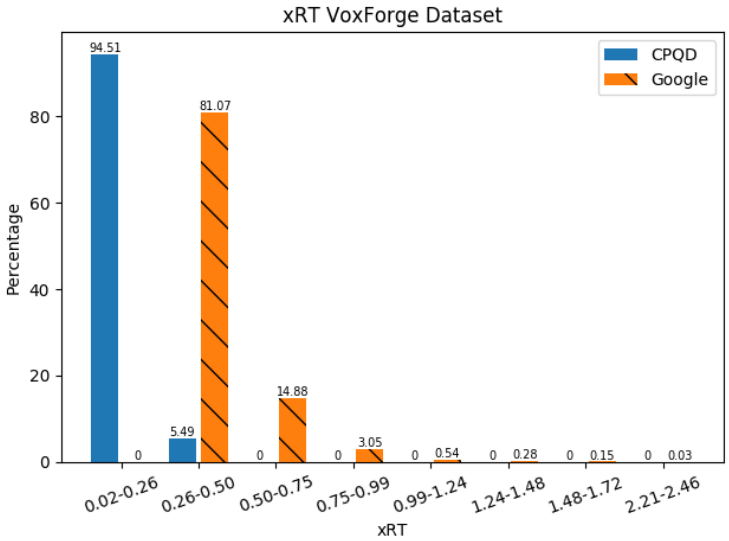

Figure 28 – Google vs CPqD xRD on VoxForge dataset.

.

sentences (i.e percent of dataset) the ASR presented a xRD among the start and ending value of the bin such bar is located.

Similar real time factor scores are observed for the CPqD ASR if compared the results in both datasets. For both ASRs a higher portion of the recognitions took more than 36% of the audio duration for the Laps-Benchmark dataset than for the VoxForge dataset. However in some cases the Google ASR took almost 2 times the duration of the audio file to output a recognition which is a lower score than that presented with the VoxForge dataset. In any case the CPqD API took more than the half of the time of the audio file to output a recognition.

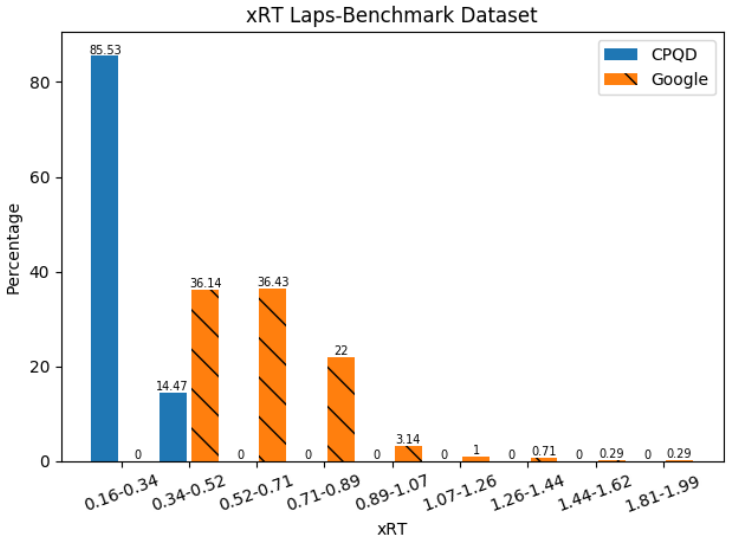

Figure 29 – Google vs CPqD xRD on Laps-Benchmark dataset.

.

Figure 30 presents the distribution of sentences by cross-entropy scores presented by both ASRs when applied on the VoxForge dataset. Each bar represents how many sentences (i.e percent of dataset) the ASR presented a xRD among the start and ending value of the bin such bar is located.

As observed in Figure 30 both ASRs presented a similar certainty on recognition of audio files in the VoxForge dataset. From Figure 31, in which is presented the cross-entropy of both ASRs on the Laps-Benchmark dataset, it is observed a higher certainty for both ASRs recognizing files in the Laps- Benchmark dataset. Such results are related to the fact that more instances and variety of conditions are present in the audio files in the VoxForge dataset than in the Laps-Benchmark dataset.

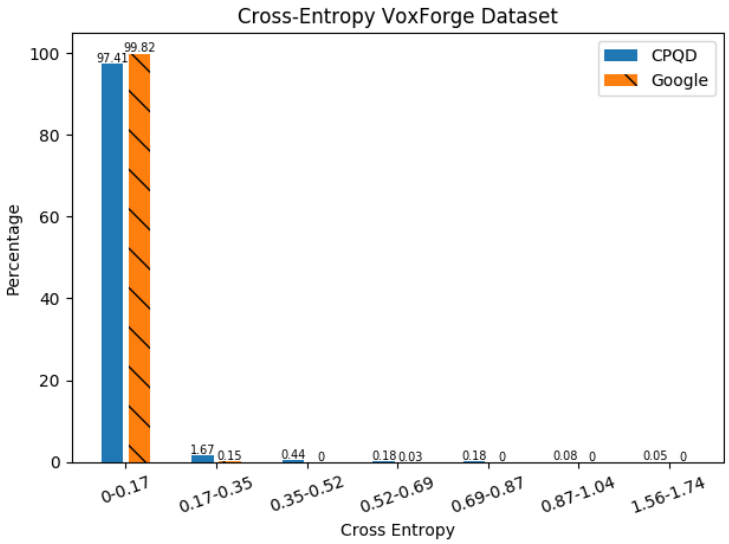

Figure 30 – Google vs CPqD Cross-entropy on VoxForge dataset.

.

5.5 Remarks

In Table 3 is shown the mean score presented by each ASR for all the 4590 audio files employed in the test. As can be seen for all cases the Google ASR presented better results except for the real time factor which mean score represents that in average using such API would take almost the half of the spoken time to recognize a given sentence. The CPqD ASR took less time to response than the Google ASR, however it should be considered that the server that provide the API of the CPqD ASR and the client application are executed in the same network on the other hand to consume the Google API a request have to pass several more nodes in the internet. Given the conditions in which the proposed system intends to function, in which a higher delay in subtitles generation would difficult understanding given the effect on synchronization, it is recommended a faster ASR.

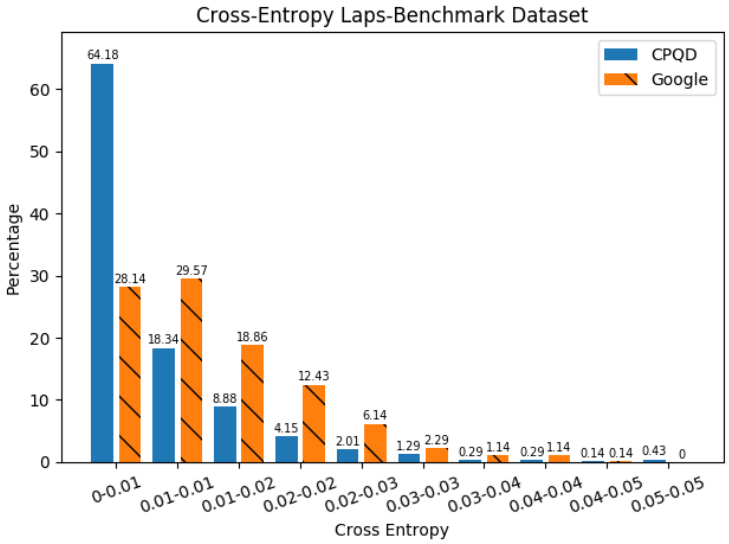

Figure 31 – Google vs CPqD Cross-entropy on Laps-Benchmark dataset.

.

ASR	WER	WRR	xRT	Hp(T)
Google	0.0818	0.92	0.46	0.017
CPqD	0.13	0.86	0.16	0.019

Table 3 – Average score of each metric by ASR.

The output of the ASR is inputed to another module intended to correct such recognized text by finding a similar sentence in the script of the play. Given the score of WRR presented by each ASR and considering that few to none improvisation is performed by an actor using the solution, it is recommended that the module for sentence correction is able to find a sentence in the script receiving as input at least 80% of the sentence's original content. However a higher flexibility can be attained dependent on the capacity of such system to recognize semantic similarity.

Even though a broad spectrum of audio duration and sentences sizes is provided by the datasets employed, which also can be spotted in the patterns expressed on metrics scores, further studies can be performed to address closer scenarios to that in which the solution is to be employed. Also a technical characterization of the audio files is recommended that permits establish comparisons among different scenarios and conditions.

6 Experiments and Results related to Speech Correction through Semantic Similarity

In the proposed method, once obtained the ASR output for an Actor speech, a module for Speech Correction is used to identify in the play script a sentence that corresponds with what the Actor said, considering also the instant in the play in which the Actor expressed such speech.

A simple approach for play's script sentence identification would be comparing, by partial or exact match, the sentence outputted by the ASR, with each sentence in the play, selecting that with a higher percent of word similarity (i.e, higher quantity of equal words in the same position). However, actors usually interpret unexpectedly, improvising or confusing the sentence that should be spoken, on several moments throughout the play, which affects sentence selection by wordwise matching.

It is expected that in most of cases the actor speaks a sentence that is semantically similar with the corresponding sentence in the script. Is for that reason that a suitable approach for Speech Correction is Semantic-Similarity-based sentence matching.

Figure 32 presents the architecture for Speech Correction (SC) considered in the proposed solution to be evaluated in this chapter. The audio of the actor's speech is submitted to an ASR in order to obtain a text of what was spoken. Such text is used as hypothesis to be compared for semantic similarity with the set of sentences in a play's script. The module for Semantic Similarity comparison receives a pair of sentences, to be compared, and outputs a score between 0 and 1, the greater the score signifies higher similarity. Finally the SC module outputs the sentence for which, when compared with the output of the ASR by the Semantic Similarity module, outputted the highest score.

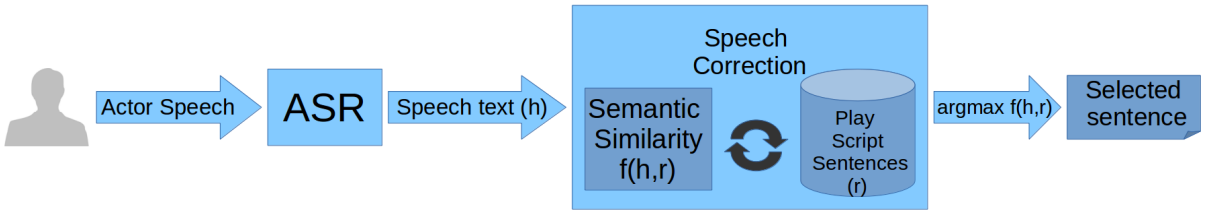

Figure 32 – Architecture for Semantic-Similarity-based Speech Correction.

The module for Semantic Similarity evaluated in this work is implemented through the python framework for natural language processing *Spacy*[1]. It integrates an statistical model that follows the architecture of a Word Embeddings by Word2Vec trained using the Universal

[1] https://spacy.io/

Dependencies[2](37) and WikiNER[3](38) Corpuses.

6.1 Experiment setup

Figure 33 presents the two setup configurations used in this work to evaluate the module of Semantic Similarity. In both configurations, pairs of sentences (h, r) are submitted to the module in order to obtain a score of similarity given by the module for such pair.

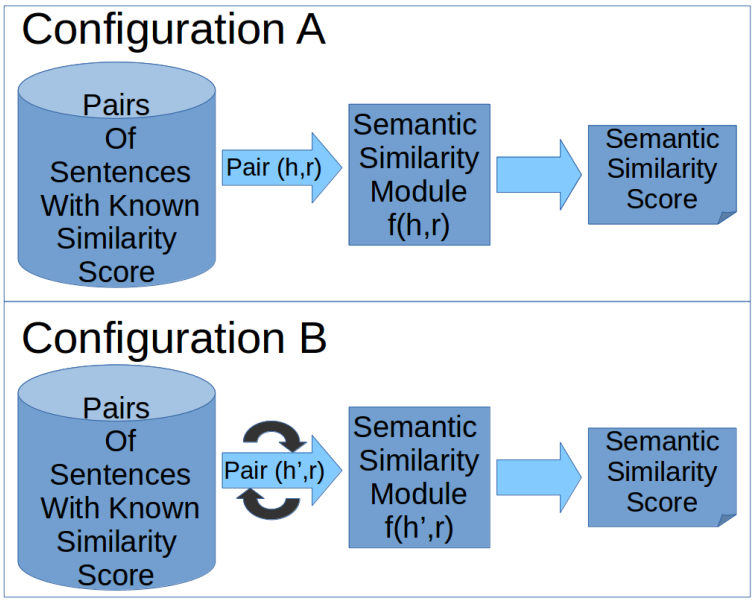

Figure 33 – Experimental setup configurations.

As the writing of this work, Speech Correction is performed employing parts of speech, i.e, as soon as the ASR returns a partial recognition of at least one word, a matching execution is performed between such output and the expected next play script sentences considering the moment of the play performance. Such process is performed in order to perform a rapid recognition, considering that the Actor will speak words that are present in the corresponding script sentence.

Configuration B presented in Figure 33 was thought to evaluate such functioning of sentence selection by partial speech recognition. For each pair in the batch, it is made a submission for each word in the hypothesis (h). Given a pair (h, r) in which the size of $h = N$ there are made N submissions h' in which are included the words of previous submissions in the same order that appears in h.

For this evaluation are considered sentences outputted by the ASRs of Google and CPqD. In chapter 5 it was addressed how such ASRs are used in order to obtain the sentences.

[2] http://universaldependencies.org/
[3] https://corpus.byu.edu/wiki/

6.2 Datasets

In this evaluation was used a dataset made of audio files recorded in an actual play with similar conditions in which the solution is to be used generally. The sentences in such audios were spoken by real Actors in a presentation of the play "O arquiteto e o Imperador da Assiria" and were manually transcribed to text files by four humans. It was extracted 459 sentences from 29.3 minutes of audio.

Five datasets were used in this evaluation, which are presented in Table 4. The datasets A, B, C and D are made of pairs of sentences in which the hypothesis is the output of an ASR and the reference is the expected output of the ASR (i.e, exact sequence of words spoken in the audio). Datasets A and B corresponds to output of Google ASR for the VoxForge and the Real Play audio sets. Datasets C and D corresponds to output of CPqD ASR for the VoxForge and the Real Play audio sets respectively. Dataset E corresponds to manually selected pairs of sentences in Brazilian Portuguese annotated by 36 human inspectors following well established guidelines, more details in next subsection.

Name	Description	Pairs	Similarity levels
A	Google ASR on VoxForge	3898	1
B	Google ASR on Real Play Audio	690	1
C	CPqD ASR on VoxForge	3898	1
D	CPqD ASR on Real Play Audio	690	1
E	ASSIN	5000	5

Table 4 – Datasets characteristics.

6.2.1 The ASSIN dataset

ASSIN stands for "Avaliação de Similaridade Semântica e Inferência Textual", Portuguese for Evaluation of Semantic Similarity and Textual Inference. It corresponds to an evaluation exercises performed during PROPOR 2016, a biennial event hosted in Brazil and in Portugal. PROPOR[4] (International Conference on the Computational Processing of Portuguese) is the main scientific meeting in the area of language and speech technologies for the Portuguese language and on the basic and applied research issues related to this language.

As a result of the ASSIN (100) exercise it was constructed the first annotated corpus of Brazilian and European Portuguese for semantic similarity and textual inference. It was compiled sentences of real texts taken from Google News in groups of different topics. Vector spaces were obtained for each word, following the approach of Turney and Pantel (101). Such vectorial models where employed to automatically select pair of sentences of different similarities. Later

[4] http://propor2016.di.fc.ul.pt/

on such pairs were filtered by four humans which defined the actual score of semantic similarity obtained they considered for each pair.

Based on a previous work that also constructed a dataset, named SICK (102), for the same tasks, it was defined five levels of semantic similarity. As the authors recognize, such measures are considered subjective and they did not get to construct an exact definition for each level of similarity, however since the pairs were analyzed by 36 humans following well defined guidelines, in this work the dataset is considered reliable. Table 5 presents the guidelines followed for pair semantic similarity score assignment and the Table 6 shows a sentence pair example of each similarity level (7).

Level	Description
1.	Sentences are completely different. It is possible that both sentences refers to the same fact, however that cannot be identified if both instances are analyzed isolated, i.e without context.
2.	Sentences refers to different facts and are not similar syntactically but fit in the same topic (a football game, voting, accidents, product marketing)
3.	Sentences are similar syntactically and can refer to the same fact.
4.	Sentence's content is quite similar but one or both, have some exclusive information. Differences among sentences can be related to the mention of a date, place, dissimilar quantities, objects or subjects.
5.	Sentences have the exact same meaning, with a minimum difference (e.g and adjective that does not alters meaning)

Table 5 – Similarity levels description.

A total of 36 annotators were trained using 18 pairs examples. Each pair is annotated by four randomly selected annotator. Later on, each pair received as score the mean of the scores given by the annotators, existing an average standard deviation of 0.49, which means low data dispersion, given the scale adopted for the levels of similarity.

Table 7 presents the quantities of pairs by similarity level. Most pairs exists with scores 2 and 3. The mean of the scores is 3.05.

Figure 34 shows the mean syntactic distance among pair's reference and hypothesis by similarity level. As expected distance is lower for higher similarity scores. For most pairs the distance is above 0.7, i.e almost three quarters of the words in the reference does not exist in the hypothesis.

Since the module for Semantic Similarity outputs values among 0 and 1, the scores in the ASSIN dataset were normalized from it scale of 5 levels to the same scale of the module. Table 8 presents the scores in the ASSIN dataset and its corresponding values after normalization.

Level	Sentence pair example
1	Mas esta é a primeira vez que um chefe da Igreja Católica usa a palavra em público. A Alemanha reconheceu ontem pela primeira vez o genocídio armênio.
2	Como era esperado, o primeiro tempo foi marcado pelo equilíbrio. No segundo tempo, o panorama da partida não mudou.
3	Houve pelo menos sete mortos, entre os quais um cidadão moçambicano, e 300 pessoas foram detidas. Mais de 300 pessoas foram detidas por participar de atos de vandalismo.
4	A organização criminosa é formada por diversos empresários e por um deputado estadual. Segundo a investigação, diversos empresários e um deputado estadual integram o grupo.
5	Outros 8.869 fizeram a quadra e ganharão R$ 356,43 cada um. Na quadra 8.869 apostadores acertaram, o prêmio é de R$ 356,43 para cada.

Table 6 – Examples of pairs by similarity level (7).

Level	Quantity of pairs
4,0 - 5,00	1.074
3,0 - 3,75	1.591
2,0 - 2,75	1.986
1,0 - 1,75	349
Total	5000

Table 7 – Quantity of pairs by similarity level

ASSIN Level	Normalized Score
4,0 - 5,00	0,8 - 1
3,0 - 3,75	0,6 - 0,75
2,0 - 2,75	0,4 - 0,55
1,0 - 1,75	0,2 - 0,35

Table 8 – ASSIN Similarity Levels normalized.

6.2.2 Datasets obtained by submitting the VoxForge dataset to CPqD and Google Automatic Speech Recognition

The audio files in the VoxForge dataset[5] are obtained from an Open Source initiative with the same name, created with the intention of collecting several audio recordings of people

[5] Available: http://www.repository.voxforge1.org/downloads/pt/Trunk/Audio/ (14/06/2018)

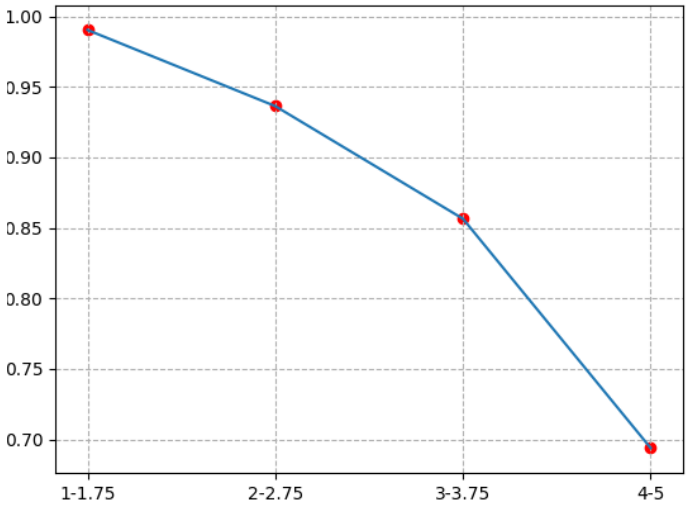

Figure 34 – Mean distance among pair by similarity level on ASSIN dataset.

worldwide, speaking in its native language, to be employed on the training of ASR engines. Such audio files were submitted to Google and CPqD ASRs and its results collected. With the output of the ASR for each sentence was created a pair in which the reference is the expected output of the ASR and the hypothesis is the actual text outputted.

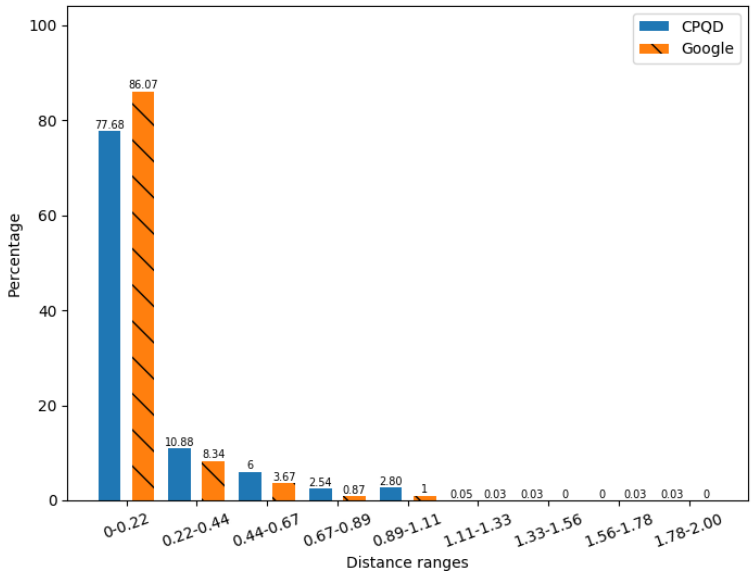

Figure 35 – Distribution of pair's sentence distance scores by ASR on the VoxForge dataset.

Figure 35 presents the distribution of sentences by distance score presented by both ASRs when applied on the VoxForge dataset. In subsection 6.3.3 of section 6.3, about metrics, it is explained how such distance score is calculated (also see Word Error Rate in (99)). Each bar represents how many sentences (i.e percent of dataset) the ASR presented a distance among the start and ending value of the bin such bar is located.

Most sentences exists with a syntactic similarity distance below 0.2, which means that for most cases at least 80% percent of the expected sentence was correctly outputted by the ASR. However, there are 450 pairs that which sentences are completely different and no information was lifted on how it affected the meaning of the sentence. Output of the module of Semantic Similarity is analyzed for each distance score.

6.2.3 Real Play dataset

The Real Play dataset was constructed through manually labeling of one real play execution audio recording, captured from the *Imperador* performer, so higher quality (and better results are expected) for the sentences spoken by such actor. Three annotators listened to approximately 10 minutes each, of different instants in the play, and annotated what was spoken by each actor.

Using the play's script as guide, the annotators associated the actor speech with each script line. The original audio recording was divided in cuts of different sizes according with the association among each line and the actor speech. For each of such cuts it was collected: *a)* Time lapse (i.e start and ending timestamps), *b)* text string of actor speech, *c)* script line associated with actor speech (i.e based on annotator judge), *d)* remark of significant event associated with actor speech (e.g if the actor improvised or introduced a sentence completely different with any on the script). Finally it was lifted a total of 459 sentences from 29.3 minutes of audio.

It was constructed pairs containing a line of actor speech lifted in annotation along with the expected to be spoken sentence in the play's script. Ideally such pairs should be equal syntactically, however in practice it mostly does not occur, due to actor improvisations. Figure 36 presents the syntactic distance among pair's sentences for each of the 459 lifted annotations. Ideally the points in Figure 36 should be grouped near 0, however a high dispersion exists in data. As seen in the figure, in the studied play, in average it was experienced a sentence modification of about 40% (i.e actors modified, by improvising, original script's sentences by 40% in average).

Table 9 presents some cases in which the distance is greater or equal to 2. It is a non exhaustive list of sentences in which it is considered that the actor introduced significant modifications on original sentence by means of improvisation. Actors spoken a total of 102 sentences as is when compared with script (i.e distance is equal 0), it represents 22% of all sentences lifted.

6.2.4 Considerations about the datasets

Regarding the dataset of ASSIN, even though sentences in the higher semantic similarity level are mostly completely different syntactically, it posses the same meaning. It is expected to occur also in real plays executions, in which actors improvise by changing most words in the original script. Such reasoning permits conclude that the ASSIN dataset is highly reliable for

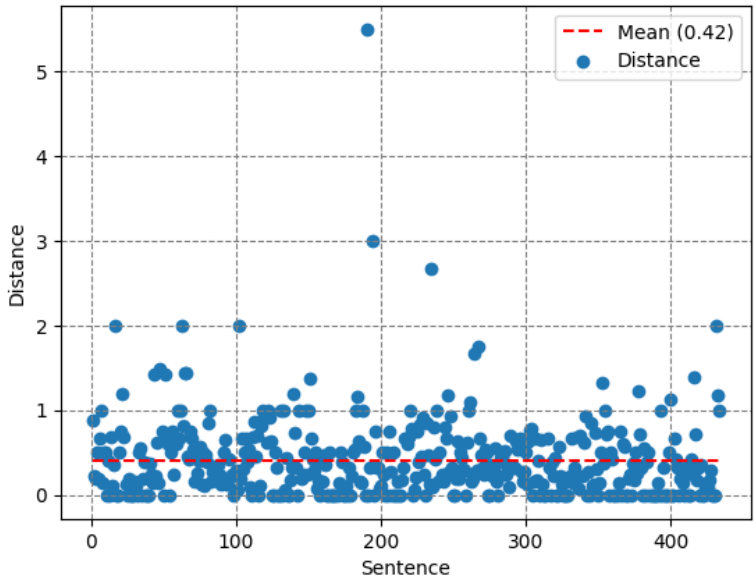

Figure 36 – Syntactic Distance among human extracted sentence and associated script quote on Real Play dataset.

Script Quote	Sentence Spoken	Distance
Charutos, charutos!	Ha ha ha ha ha ha. Charu. Ha ha ha ha	5.5
Bom, me lembrarei dessa atriz, como se chama?	Hã, pensando naquela atriz gostosa. Que fez aquela novela sabe, aquela novela, você sabe aquela novela?	2
Adeus	Adeus para sempre	2
Você disse	Você disse que eu podia chamar de você	3
Eu não estava preparado.	Não vale, não vale, você estava preparado, não vale, volta aqui	2
Sim. (*Sentence associated by correspondence considering previous and next sentences.*)	Que isso?	2
(*Sentence could not be associated with any in the play script*)	Está sim! Está Sim! Você não pode nem compreender, não pode!	3

Table 9 – Example of pairs with distance greater than equal 2 in real play dataset.

the evaluation and broadly represents general scenarios for which the proposed solution should be prepared. Different than the other datasets, the ASSIN dataset presents a broader spectrum of categories that permits evaluate the module in the edges of similarity values that are mostly expected to see in real plays scripts. Such granularity represents several scenarios in which the Speech Correction can fail if the accuracy of the Semantic Similarity does not correctly discriminate among each category.

Datasets A and C can be considered as representing a scenario in which no improvisation is made, and Speech Correction would have to deal only with modifications introduced by the ASR. In such case the quality of the ASR have a direct influence in the result of Speech Correction, however, the question arise if the module of Semantic Similarity is effective for syntactically similar pairs.

Datasets B and C represents an actual play and presents not only modifications introduced by the ASR, but actor's improvisations. Since the amount of data collected from datasets B and C cannot be considered representative of several plays, it is not intended to conclude from such datasets how often and in which magnitude actors improvisation occurs in real play executions. Actor's improvisation and characteristic of performance have particularities for each play and not necessarily exists a pattern, not even for the same play, however it is expected to lift from the output of the ASR and Semantic Similarity modules applied on such datasets, its performance for plays with similar characteristics.

6.3 Evaluation metrics

In order to evaluate the quality and using feasibility of the module for Speech Correction based on Semantic Similarity, a set of metrics are lifted from the parameters observed on the batch of submissions executed following the setup presented in section 6.1. Such metrics are presented in following subsections.

6.3.1 Semantic Similarity Mean Absolute Error

This metric corresponds to the deviation of the Semantic Similarity module's output (y_i) compared with the expected score (x_i). Since it is previously known the score for each pair of sentences it is possible to establish such comparison. In equation (6.1) is formally presented how this metric is calculated.

$$MAE = \frac{\sum_{i=1}^{N} |y_i - x_i|}{N} \tag{6.1}$$

For most results of this metric presented in following subsections, the values of N in equation 6.1 corresponds to groups of similarity. For the ASSIN dataset such groups corresponds to the defined levels of similarity. For the datasets obtained by the use of an ASR system two

approaches are considered to select such groups, the first is based in the assumptions that both the ASR output and the expected sentence are equal semantically and its similarity score is 1, the second approach groups pairs by the distance of its sentences.

6.3.2 Server response time

It is also of interest in this evaluation to lift the time it takes for the module to output a score for a given pair. In a real scenario in which the solution is to be executed it is quite important to deliver the correct subtitle in a moment not too distant from that in which the acting corresponding to such subtitle occurs, under the assumption that the user do not understand the relation of a subtitle showed in a different context than that being performed.

$$TOR = Timestamp\ of\ response - Timestamp\ of\ submission \tag{6.2}$$

Equation (6.2) formally introduce the formula used to calculate this metric. In such equation timestamp stands for the number of seconds that have elapsed since January 1, 1970 at 00:00:00 GMT (1970-01-01 00:00:00 GMT). Results for this metric are expressed in seconds.

6.3.3 Syntactic distance

It is also known as the Word Error Rate (WER) and is employed in most ASR applications as the figure of merit (99). Such metric is calculated for a pair of sentences known as the reference and hypothesis. In an ASR system evaluation, the WER is a proportion of how many errors exists on recognition against how many words exists on the reference, based on: *a)* substitutions, the words in same positions of hypothesis and reference are different; *b)* deletions, words of the reference does not appear in the hypothesis; *c)* insertions, different words than those in the reference are included in the hypothesis. In equation (6.3) is presented the formula employed to calculate the syntactic distance (d).

$$d = \frac{S + D + I}{N} \tag{6.3}$$

where: S = substitutions
D = deletions
I = insertions
N = size of reference

To compute the syntactic distance, the sequence of words in the reference and hypothesis are extracted and lowercased so no symbols or spaces in any of both are considered. To identify incorrectly recognized words an algorithm of sequence alignment is applied using as input both sequences (i.e reference and hypothesis) to find the longest contiguous matching subsequence

that contains no junk elements. Then the algorithm is applied recursively to the pieces of the sequences to the left and to the right of the matching subsequence. To such end, in this study was employed the implementation provided by the python library *edit_distance*[6].

	Position										
Ref/Hyp	0	1	2	3	4	5	6	7	8	9	10
REF:	se	conseguirmos	UMA	PARTE	DO	que	ELE	alcançou	será	**	DEMAIS
HYP:	se	conseguirmos	O	APORTE	DE	que	***	alcançou	será	DE	MORTE

Table 10 – Example of reference hypothesis alignment.

Table 10 presents an example of output of the algorithm for sequence alignment. Lower cased words represents matches and upper cased represents words that did not matched. When asterisks appears in the reference it means that in this position of the hypothesis was inserted a word, in the example was inserted the word *DE* in position 9. When asterisks appears in the hypothesis it means that the word on that position was deleted, in the example the word *ELE* was deleted in position 6. When upper cased words appear in the same position on reference and hypothesis it means that a substitution occurred, in the example was substituted the words from position 2 to 4 and in position 10. For the example presented in Table 10 the value of d is:

$$d = \frac{1(ins) + 1(dels) + 4(subs)}{10\ (ref\ size)} = \mathbf{0.6}$$

6.4 Results

In this section are shown the results of submitting the datasets presented in section 6.2 to the modules for Speech Recognition and Semantic Similarity.

6.4.1 Server response time

Since it was experienced similar response values for all submissions done for all datasets all results were compiled in a single section. Table 11 presents the average response times of the semantic similarity module server on each dataset.

It is worth noting that such response time is significantly lower if the statistical model, employed in the semantic similarity scoring, is loaded when the server is initiated and not on the processing of each submission (i.e, on each request made to server). However, such approach does not permit update the model on the fly and the server should be restarted in such case.

[6] Available: https://github.com/belambert/edit-distance (14/06/2018)

Name	Dataset Description	Mean Response Time (Secs.)	Mean Absolute Deviation
A	Google ASR on VoxForge	0.15	0.05
B	Google ASR on Real Play Audio	0.13	0.02
C	CPqD ASR on VoxForge	0.14	0.01
D	CPqD ASR on Real Play Audio	0.15	0.03
E	ASSIN	0.13	0.02

Table 11 – Server response time by dataset.

6.4.2 Semantic Similarity on the ASSIN dataset

Figure 37 presents the Mean Absolute Error (MAE) by normalized similarity level of the ASSIN dataset, presented by the module of Semantic Similarity (SS). The red vertical lines represent the Absolute Error score's dispersion for each similarity level or what is the same the Average Absolute Deviation by similarity level. Both measures express how the output of the SS module deviates from the expected, the first is the average of the distances from the expected to the outputted score and the second the average distance from all the error values to the mean.

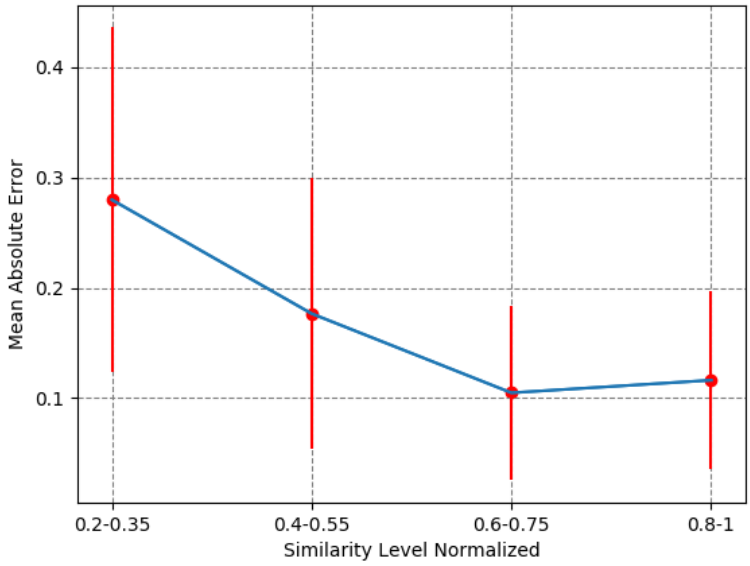

Figure 37 – Mean Absolute Error by similarity level on the ASSIN dataset.

The SS module experienced higher accuracy for most similar pairs, i.e pairs in the similarity level score 5. Error values dispersion is lower for higher similarity levels, which means also more certainty and centrality of accuracy for such pairs. However the error for pairs made of semantically unrelated sentences in some cases induced to the incorrect income that sentences in such pairs have the same meaning since it mislabeled pairs into the category of quite similar sentences.

If considered that most pairs in each level have as score that in the center of its scores range, for all levels the error margin surpasses each category of similarity. It means that if the set of sentences to be compared by the Speech Correction module are semantically close, a high chance exists that such selected sentence is an incorrect one.

Out of 349 instances in the group of pairs with least similar sentences (i.e, score of normalized similarity below 0,35), 35 samples was put in the group of most similar. On the other hand only 3 out of 1.074 instances in the group of most similar sentences (i.e, similarity score normalized above 0,8), were put in the group of least similar. Table 12 presents examples of such cases.

H/R	Text	y	y'
R	Em maio, endividamentos interno e externo somavam R$ 2,49 trilhões.	0.85	0.28
H	Em maio, a dívida estava avaliada em 2,49 trilhões de reais.		
R	A previsão para a taxa de câmbio em 2015 ficou em R$ 3,20.	0.25	0.8
H	Para 2016, a previsão de superávit comercial permaneceu em US$ 9,95 bilhões.		
R	De acordo com o conservacionista, um grupo caçava a noite, quando avistou Cecil.	0.8	0.2
H	Eles saíram para caçar à noite e encontraram Cecil.		
R	A indústria extrativa exibe comportamento totalmente distinto do setor de transformação.	0.2	0.8
H	O início de ano exibe comportamento de queda para além da volatilidade.		
R	A mudança nas taxas entra em vigor em 1º de outubro e não atinge os contratos do Minha Casa Minha Vida.	0.8	0.31
H	As novas taxas vão passar a vigorar em 1º de outubro.		

Table 12 – Examples of sentences put in opposite extreme.

In Table 12 the first column (H/R) has value R for rows which text in the second column represents the reference of the pair and H for the hypothesis. The column with header y represents the expected value (i.e manually assigned score for pair) and y' represents the module outputted score.

6.4.2.1 Similarity on partial hypothesis of the ASSIN dataset

Figure 38 presents the Mean Absolute Error by percent of the hypothesis sent to the module of Semantic Similarity on the ASSIN dataset. As expected the error approximates to the mean error for the overall dataset as a higher part of the hypothesis is submitted, which is reached when 80% of the hypothesis is submitted. If considered that the distance among similarity levels

is approximately 0.2, from over, sending less than a half of the hypothesis sentence would put any pair in the incorrect category.

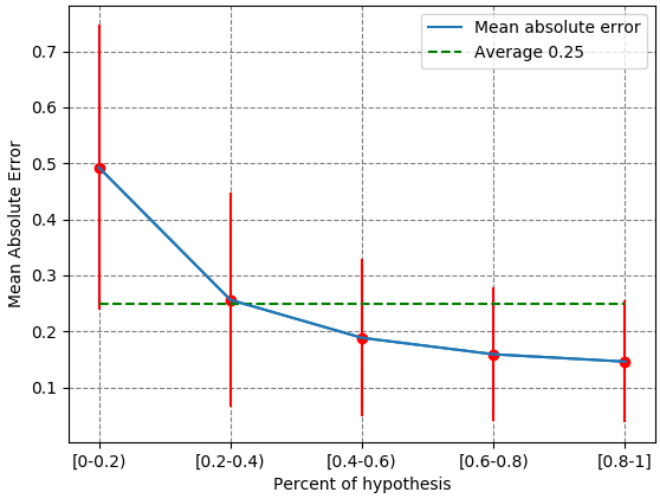

Figure 38 – Mean Absolute Error by percent of hypothesis on the ASSIN dataset.

6.4.3 Semantic Similarity on VoxForge dataset, Google vs CPqD Automatic Speech Recognition

Figure 39 presents the Mean Absolute Error of Semantic Similarity scoring on the output of Google and CPqD ASRs when applied on the VoxForge dataset, by distance among sentences of the pair. It is worth recalling from a previous section that in this experiment the expected text to be outputted by the ASR is taken as the reference and the hypothesis is the actual ASR's outputted text. Also the expected Similarity score to be outputted by the Semantic Similarity module is of value 1.

The error is lower for pairs the ASR recognized with higher accuracy. For all pair's sentences distance values, better results are observed for the output of Google. Pairs with distance among sentences lower than 0,44 (i.e almost the half of sentence is different), were recognized with similar error for the output of both ASRs.

Figure 40 presents the Mean Absolute Error of Semantic Similarity scoring on the output of Google and CPqD ASRs when applied on the VoxForge dataset, by percent of hypothesis submitted. As more words of the hypothesis are sent, lower is the error, as expected.

This experiment represents the case in which no improvisation is made by actors, since the hypothesis is expected to be equal to the reference. The results presented permits evaluate the influence of the ASR recognition accuracy on the accuracy of the module of Semantic Similarity. As expected when less information on the pairs is presented to the module, worst results are obtained. In all cases the best results were obtained on the output of the Google ASR.

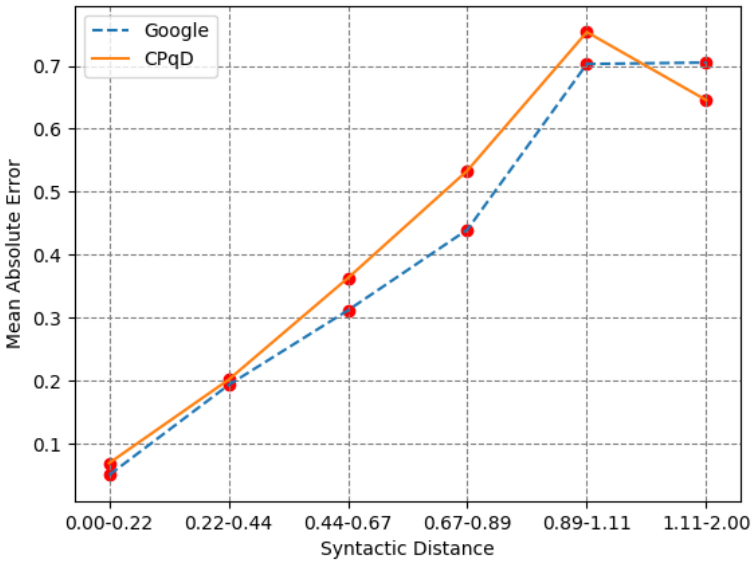

Figure 39 – Similarity Mean Absolute Error for outputs of Google vs CPqD on VoxForge dataset by distance among hypothesis and reference.

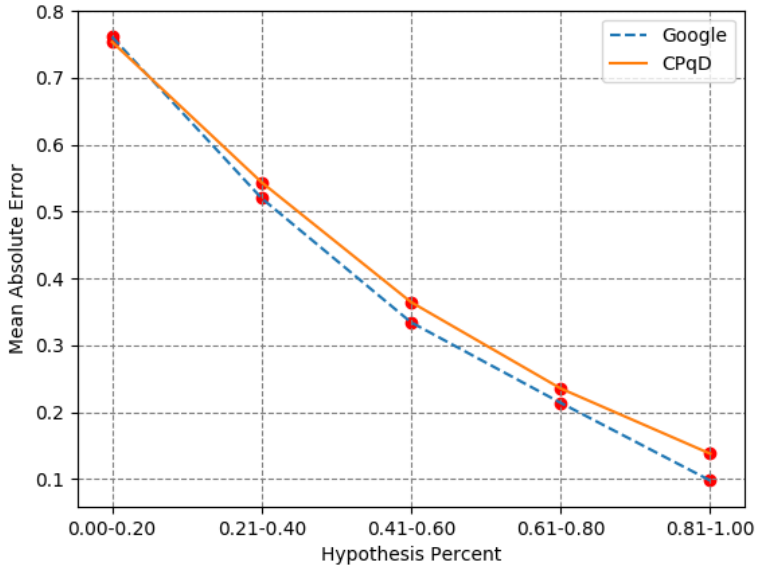

Figure 40 – Similarity Mean Absolute Error for outputs of Google vs CPqD on VoxForge dataset by percent of hypothesis submitted.

Sending partials of the hypothesis permits reduce the time taken by the system to output a subtitle, however as more of the ASR outputted sentence, that corresponds with the script sentence being performed by the actor, is presented to the module for Semantic Similarity, better results are obtained. In a real play, analyzing the syntactic distance among sentences in the set of those compared to select the correct subtitle, and the percent of improvisation generally observed for such lapse of the play, is possible to estimate the percent of hypothesis necessary enough to select the correct sentence with a defined error margin.

6.4.4 Semantic Similarity and Automatic Speech Recognition output (Google vs CPqD) on Real Play dataset

By submitting all audio pieces in the Real Play dataset to each ASR was obtained the recognized text for the speech on each piece. Three groups of sentences pairs were considered: *a)* sentences spoken by actors (manually annotated by human) with its corresponding in the play script, *b)* sentences spoken by actors with its corresponding outputted by the ASR, *c)* output of the ASR with its corresponding sentence in play script. Syntactic and semantic scores were lifted in order to observe the accuracy of the ASRs and Semantic Similarity modules. For groups *a)* and *b)* it is expected a syntactic distance of 0 and semantic similarity of 1.

Each point in Figure 41 represents the Absolute Error of the Semantic Similarity module when applied to a pair of the group *a)* (i.e sentences spoken by actors with its corresponding in the play script). In Figure 41 are not plotted any point for sentences that could not be related semantically or syntactically with any in the play script. Ideally such values should be 0 since it was judged by a human that the spoken sentence have the same meaning as its associated sentence in the play script, even though it high syntactic distance as presented in Figure 36. However as seen in Figure 41 there is a high dispersion in data, existing outputs of different error scores. When applied on group *a)* the module for Semantic Similarity presented a Mean Absolute Error similar to that presented when applied on the ASSIN dataset (i.e approximately 0.17).

Table 13 presents pairs of sentences in the group *a)* for which the module of Semantic Similarity presented an unexpected score. Such examples are highlighted as significant since sentences are equal or mostly equal syntactically, thus is expected the maximum score of semantic similarity (i.e 1).

It is observed, in the examples of Table 13, that the quality of the output given by module presented an error in some cases of 30%. Such results, subjectively, can be considered poor, since it means that even when the actor did not performed any improvisation the system was not able to assign a correct score.

Figure 42 presents the scores of syntactic distance (d) and semantic similarity absolute error by group of pairs and ASR, considering all sentences (i.e without distinction by actor). From

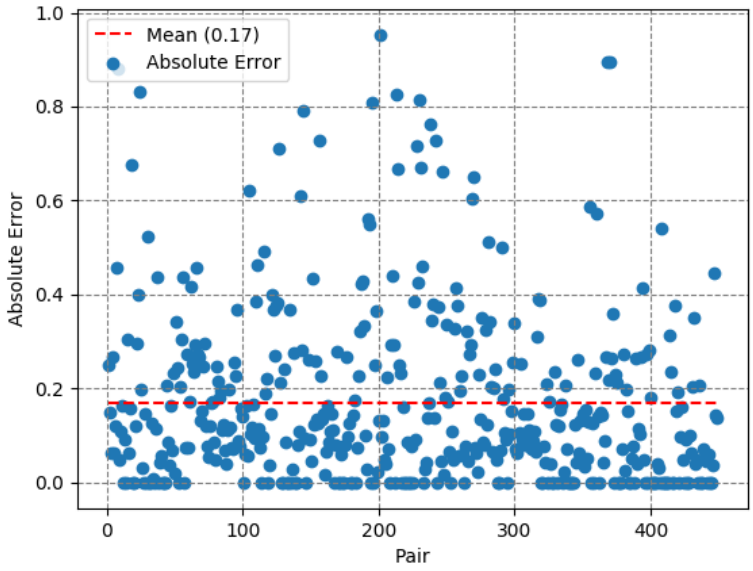

Figure 41 – Semantic Similarity Absolute Error on Real Play dataset.

Hypothesis	Reference	Semantic Similarity
Começando pelos senhores mesmos.	começando pelos senhores mesmos.	.964
Vou buscar o binócolo	Vou buscar os binóculos.	.687
Quem ? Quem ?	Quem, quem?	.968
O quê que tá acontecendo ?	O que está acontecendo?	.819
E a odiava ?	A odiava?	.887
A senhora é esposa do acusado ?	A senhora é a esposa do acusado?	.913
Silêncio, que entre a primeira testemunha	Silêncio. Que entre a primeira testemunha.	.956
Sim	Sim, senhor.	.731

Table 13 – Example of pairs with evident similarity that were mislabeled by Semantic Similarity Module

this figure it is observed that semantic similarity was less affected than syntactic, as expected. For both ASRs, when applied on the group of pairs *a)* the error in recognition surpasses 40%, significantly worst when compared with the output of such ASRs on the VoxForge dataset. It occurs given the characteristics of the speech in the Real Play dataset, in which the actors mostly shout and laugh.

Figure 43 presents the scores of syntactic distance (d) and semantic similarity absolute error by group of pairs and ASR, considering only sentences spoken by the *Arquiteto* performer. The error presented by both ASRs in such audios is high and similar, it occurs due its poor quality, since it was captured from the microphone in the other performer and was mostly heard in a lower volume. Syntactically, the output of the ASR was almost completely different with the sentence in the script given the improvisation of actor and the error of ASR. The output of the

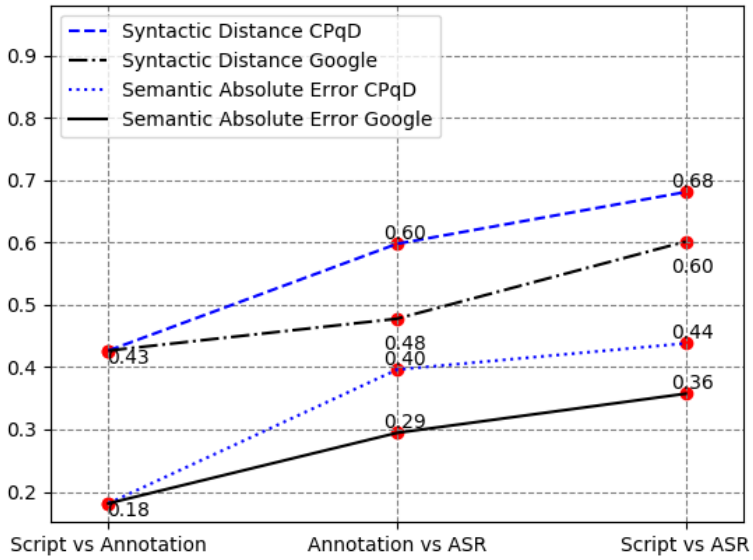

Figure 42 – Syntactic and Semantic error introduced by each ASR.

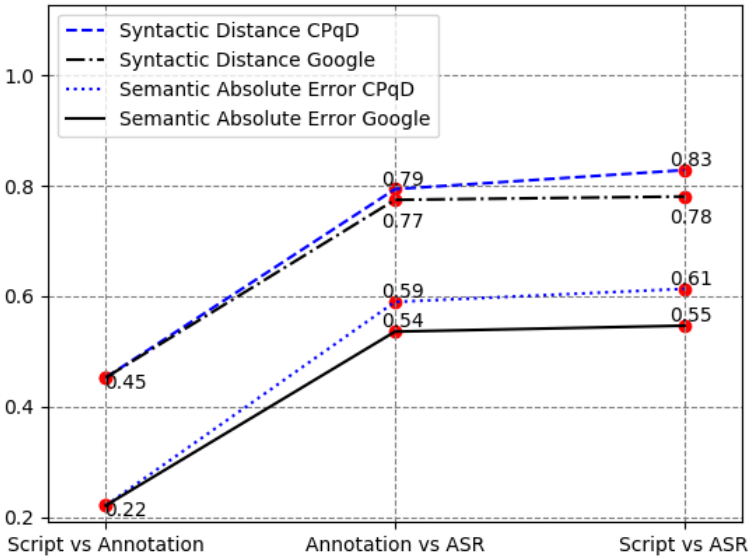

Figure 43 – Syntactic and Semantic error introduced by each ASR on *Arquitecto* sentences only.

ASR affected negatively the output of the Semantic Similarity module in about a 40%.

Figure 44 presents the scores of syntactic distance (d) and semantic similarity absolute error by group of pairs and ASR, considering only sentences spoken by the *Imperador* performer. According to the Semantic Similarity module output and the syntactic similarity observed in the pairs of group *a)*, the amount of improvisation for both actors was similar. Since the microphone was closer to *Imperador* performing actor the audio presented a higher quality and it influenced in the output of ASR when compared with the audio of *Arquiteto*, the difference in error is of a 30% approximately. However the output is still worst when compared with output on the VoxForge dataset, due to characteristics of harsh speech.

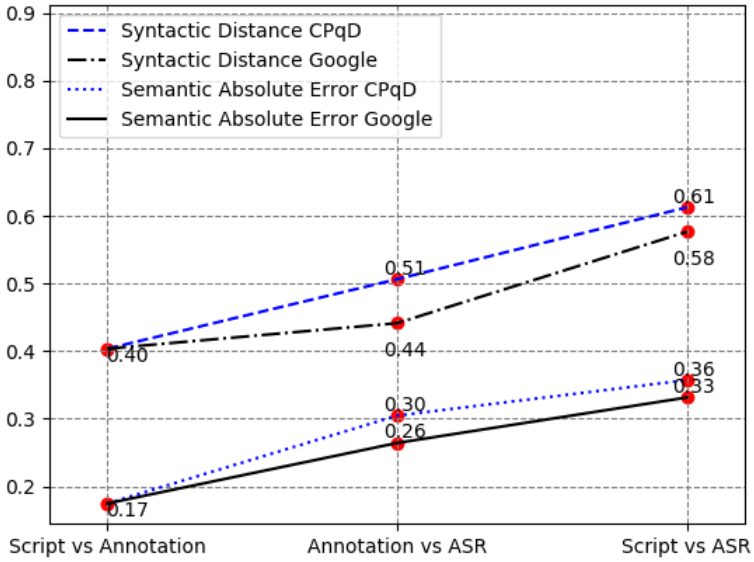

Figure 44 – Syntactic and Semantic error introduced by each ASR on *Imperador* sentences only.

6.5 Remarks

In this evaluation was assessed accuracy and performance of the modules for Automatic Speech Recognition and Semantic Similarity integrated. Such modules was applied to five datasets that represents different scenarios that the solution should be prepared to deal with, which can be described as: *a)* Actors syntactically modify the play script by improvisation and ASR did not introduced any modification, *b)* Actors and ASR syntactically modify the play script by improvisation and ASR error on recognition, *c)* Actor did not introduced any modification but ASR did.

The values lifted in this evaluation can be used as baseline for comparison on future enhanced versions of the solution. Also it can be employed to define a threshold of similarity when selecting the correct sentence in the process of Speech Correction, considering the difference among sentences in the script and the deficiencies of the modules evaluated.

The ASRs performed worst for the Real Play dataset than for the VoxForge dataset. It occurs due to the characteristics of the speech in the Real Play dataset, in which the actors speech is not plain but rather laughing and shouting. Such not ordinary way of speaking is usual in real plays and is expected that the system is able to perform well in such conditions. In order to obtain better results it is recommended that the model for speech recognition is trained with audio presenting the same harsh characteristics of speech.

Speech Correction based on Semantic Similarity would present less error than Syntactically based. However, the output of the Semantic Similarity module is affected by the error introduced by the ASRs. It is recommended fine tuning and customizations for each play as well as further study on several actual play executions in order to identify patterns that permit enhance the accuracy of the models for Automatic Speech Recognition and Semantic Similarity.

7 Experiments and Results related to Subtitling Method

In this chapter all finding regarding subtitling in summarized. Moreover, the followed methodology to achieve the results is described pointing out the real scenario used to better collect solid results. The main focus on this chapter is summarizes the results in users (DHH community) perspective.

7.1 Experiments and Results related to Text Subtitling

In this section is detailed the experiment to evaluate the text subtitling method from experiment setup to results. Then, to present a consideration about the system.

7.1.1 Experiment Setup

The experiment to evaluate the system proposed in this thesis focused on text captioning was divided in two parts covering two different plays.

In 2017, the first part of experiment was done from May 12th to July 23th, we conducted 10 user testing (UT) sessions in two mainstream theaters in São Paulo city, Brazil. The Figure 45 shows participants using Gear VR after the play in Fernando Torres theater, which is showed in Figure 46. Structured data and qualitative insights were collected from 43 DHH attendants over weekly performances of 'O Pai' play - 'The Father', from the original French 'Le Père'.

Continuing the experiment, the second part was executed in 2018, from April 13th to May 18th, we conducted 4 user testing (UT) sessions in one mainstream theater in São Paulo city, Brazil. Structured data and qualitative insights were collected from 10 DHH attendants over weekly performances of 'O arquiteto e o imperador da Assíria' play - 'The Architect and the Emperor of Assyria', from the original Spanish 'El arquitecto y el emperador de Asiria'. A play scene is showed in Figure 48.

This initiative was used as a marketing campaign sponsored by Samsung [1].

Figure 47 shows the cast of play 'O Pai' and summarizes the technical information about play staged during tests.

Participants were selected with aid from regional Deaf association. Before play, participants were trained by supervisors about how to use the app on the VR device (Gear VR + Galaxy

[1] https://www.youtube.com/watch?v=RRihTJxAlk0

Figure 45 – User's Test Participants and instructors.

S7 - for 'The Father' play and Galaxy S9 for 'The Architect and the Emperor of Assyria' play), eventual experiment issues and quick fixes, in case they occur.

After play, participants answered 4 structured questions about image/display, subtitle, understanding and satisfaction using Likert-scale (1 poor to 5 best):

- Subtitle: if they could read transcriptions with proper timing and readability
- Image/display: whether they could see actors and stage with desired quality
- Understanding: whether they could get the entire stream of speeches and emotions
- Satisfaction: how pleasant and rewarding it was to use the VR captioning system

7.1.2 Results

All collected quantitative data are summarized in Figure 50 for 'The Father' play and in Figure 51 for 'The Architect and the Emperor of Assyria' play.

7.1.2.1 Subtitle Evaluation

Subtitles had some dispersion on votes, but a consistent amount of these stood around good opinions. This cited dispersion was noted in the two parts of experiment. This means that subtitles may have performed well, but there is still much room for improvements. Users also suggested better synchronization and additional features to regulate caption size, adjust its placement and contrast on-screen.

Figure 46 – Deaf or hard-of-hearing people using the system described in this experiment in the play 'O Pai'.

7.1.2.2 Image/Display Evaluation

In the two parts of experiment, most of participants had neutral to bad opinion about image provided by Gear VR + Galaxy S7/S9. Thus image/display was the worst factor of all analysis, mainly due to its inability to manage light intensity and to provide the desired definition. Some users suggested the addition of features such as zoom, focal adjustments, and brightness control. Even as mobile device update, resulting in more resolution to AR simulation environment, the result kept the same. Then, to improve this item a truly AR device is recommended.

7.1.2.3 Understanding Evaluation

Overall understanding feedback were good in the first phase, but we've noted an evident decrease in this item in the second phase. Our main hypothesis to explain it is related to 'The Architect and the Emperor of Assyria' play, which by its lack of context and even of text the play lends itself to various interpretations (103). Many participants freely stated that not only the technology helped on the understanding but also that it was actually better than using professional interpreter services.

7.1.2.4 Satisfaction Evaluation

Overall satisfaction were good in two experiment parts, which means participants could follow all the play and be aware of surrounding spectators' emotions. Actually, it was the first

Writer	Florian Zeller
Director	Léo Stefanini
Translators	Carolina Gonzalez and Lenita Aghetoni
Cast	Fulvio Stefanini, Carol Gonzalez, Lara Córdulla, Carol Mariottini, Paulo Emilio Lisboa and Wilson Gomes

Figure 47 – Cast actors and Information about the play "O Pai".

time to enter in a theater for many participants, which means a more accessible place to visit and enjoy it.

7.1.2.5 Experiment Limitations

All participants were invited by Samsung (free of charge); It was noticed some observer-expectancy effect; It is still important to test the device in different plays styles.

7.1.2.6 Improvement chances

Great majority of issues seems to have an integrated solution with the adoption of lighter and unobtrusive AR glasses instead of Gear VR: from light intensity to image definition, facial expressions, inadequate rendering of stage lights, excessive brightness, head and eye strain, and excessive device weight. All seem to be easily solved by most of AR concepts available and probable to come.

Figure 48 – Scene of the 'The Architect and the Emperor of Assyria' play.

Some other issues, however, can yet be discussed for further optimization: better software validation, overheating and simultaneous use of correction glasses.

7.1.3 Remarks

Captioning method empowers DHH people communications and this work extended this convenience to live theaters with lower costs, proposing a specific application that has not been investigated so far. Systems like this are far from optimal with many unsolved challenges for the subtitle generation. In general, subtitle systems for theaters, supported by our proposed technology, were well accepted by DHH spectators. UT pointed that participants could follow the entire play with good understanding of both scene rationale and crowd emotion, with results pointing that these two components are crucial drivers for user overall satisfaction. Subtitles had good reviews with lesser complaints about subtitle synchronization; and images had bad reviews mostly because of camera and hardware limitations.

7.2 Experiments and Results related to Sign Language Subtitling

In this section is showed the experiment to evaluate the sign language subtitling method from experiment setup. Then, to summarize the results to present a consideration about the system.

Figure 49 – DHH users watching the play.

7.2.1 Experiment setup

The experiment was executed in 2018, from July 8th to Oct 5th, we conducted 7 user testing (UT) sessions in one mainstream theater in São Paulo city, Brazil. Structured data and qualitative insights were collected from 20 DHH attendants over weekly performances of 'O arquiteto e o imperador da Assíria' play - 'The Architect and the Emperor of Assyria', from the original Spanish 'El arquitecto y el emperador de Asiria'. In contrast to last experiment described in last section, in this experiment the sign language was used to subtitling instead of text.

As in last experiments, all participants were selected with aid from regional Deaf association. Before play, participants were trained by supervisors about how to use the app on the VR device (Gear VR + Galaxy S9), eventual experiment issues and quick fixes, in case they occur.

Aiming to have a reference, the same listed questions in the last section about subtitle, image/display, understanding and satisfaction were used in this experiment. The different in subtitle evaluation item is the UI changing from text to sign language subtitling. Other improvement in this experiment is the usage of semantic similarity in speech correction module, this can infer in all evaluation items. Finally, participants went into an interview that collected qualitative insights about their experience along the play.

7.2.2 Results with Sign Language Subtitling

All collected quantitative data are summarized in Figure 52.

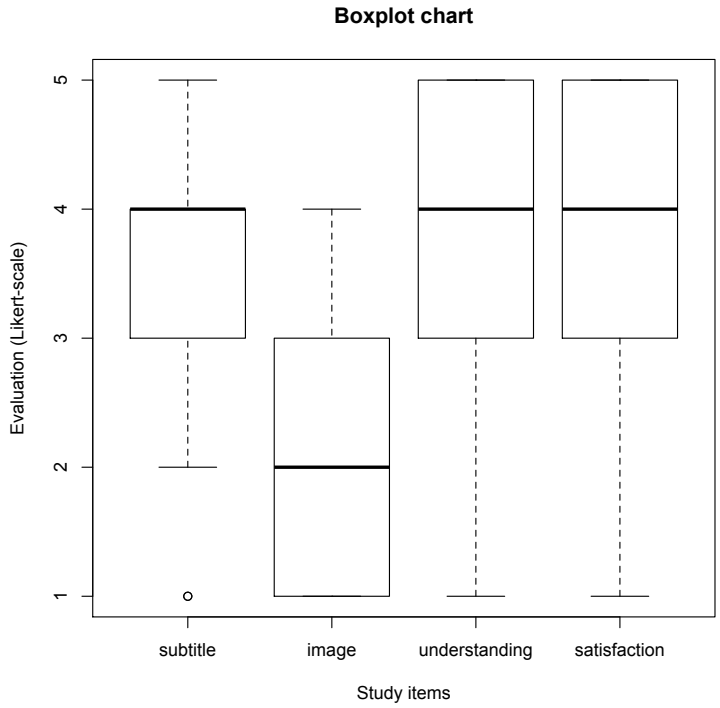

Figure 50 – Boxplot to summarized all quantitative data collected from 'The Father' play.

7.2.2.1 Subtitles (sing language subtitling) Evaluation

Subtitles using sign language had a large dispersion on votes, but a consistent amount of these stood around neutral opinions. In comparison with text subtitling the results are significantly worse. Based on that, we can figure out sign language has a lot opportunity to be improved, but cannot be ruled out. More information will be showed in qualitative feedback in the next section, which will guide us to understand user needs to better design this solution.

7.2.2.2 Image/Display Evaluation

As in the last experiments, the same evaluation was measured. Most of participants had neutral to bad opinion about image/display keeping the conclusion a truly AR device must be used.

7.2.2.3 Understanding Evaluation

A very similar evaluation based on last experiment was noted. Even though, sign language subtitling received a worse evaluation, this didn't affect strongly the play understanding.

7.2.2.4 Satisfaction Evaluation

Overall satisfaction were good in this experiment, which means participants could follow all the play. As in last experiment, it was the first time to enter in a theater for many participants.

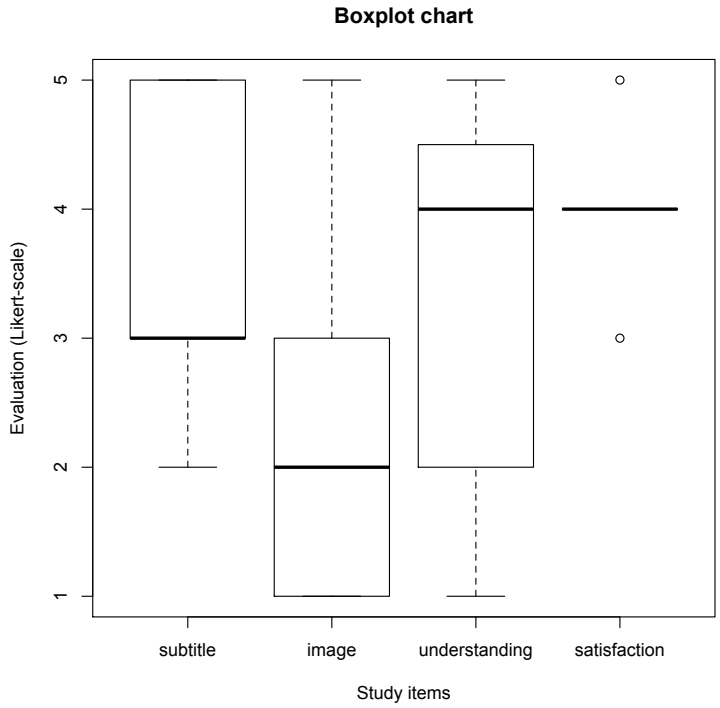

Figure 51 – Boxplot to summarized all quantitative data collected from 'The Architect and the Emperor of Assyria' play.

Mainly, participants which don't know to read and it was the first opportunity to follow a live play through sign language window.

7.2.3 Remarks

Based on user's feedback collect by interviews, some insights related to this proposed solution were found.

Regarding the sign language window, although this item was evaluated with neutral opinions, based on user's feedback, we've evidences the interpreter's performance had a significant impact on the evaluation of this item. This is fully related to this work (84), when authors' review revealed the importance of the interpreters' relationship with the content. In addition to 'ease of shifting attention', a relationship can be demonstrated via the direction of the interpreters' and their synchronized interaction with the TV content, concluded the authors. Similar results were found and the sign language translation for a play is a huge challenge. More studies about this topic are needed.

Concerning VR/AR device usage, a very interesting feedback was collected. A participant pointed out the VR device, which emulates an AR device, oppresses the impaired people confronting the inclusion perspective. Aiming to solve image issue and this, we believe an evolution of AR technology will improve this solution when AR glasses made like regular glasses will remove this excluding perception.

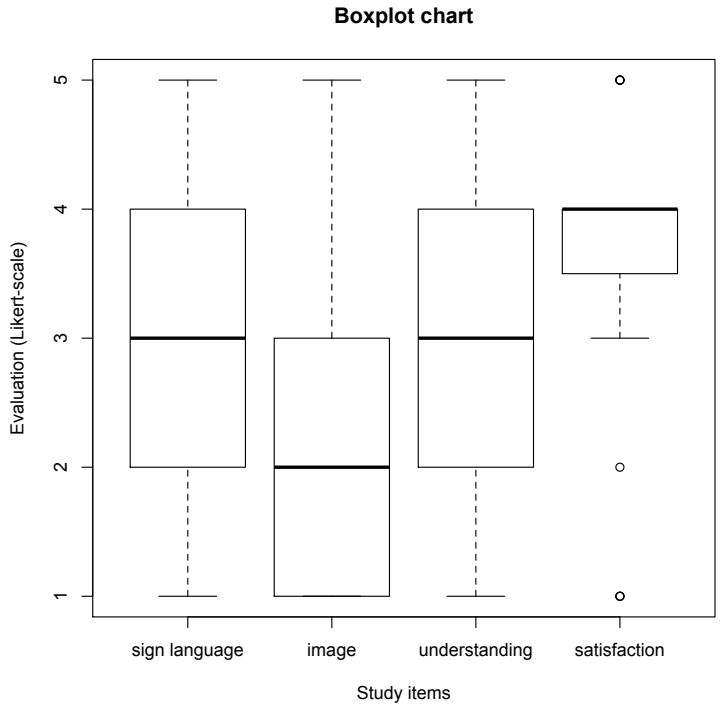

Figure 52 – Boxplot to summarized all quantitative data collected from 'The Architect and the Emperor of Assyria' play using sign language subtitling.

One important theme in captioning is whether the implementation of captions in individual sign language interpreter videos can positively affect viewers' comprehension when compared with sign language interpreter videos without captions. In this work (104), authors showed that the presence of captions positively affected their rates of comprehension, which increased by 24% among deaf viewers and 42% among hard of hearing viewers. This investigation in VR context is cited here as a next evaluation theme.

The sign language subtitling is a promising technology, but a huge effort is necessary to start a standard for displaying definition in a VR/AR device. Actually, neutral feedbacks were collected and the qualitative study pointed out to sign language subtitling interpreter acting, which must be synchronized with actors regarding time, expression and feeling demonstration.

8 Conclusion

In this thesis is presented the state of the art regarding Virtual Reality technology usage in accessibility systems. Moreover, it's summarized the results from implemented case study, focusing to improve the understanding of how DHH people can make VR environment useful as an assistive tool.

In the section 4.1 ("Text Subtitling Method to DHH People in Live Theaters"), a subtitle system for theaters supported by VR technology was designed, implemented and validated. The performed experiment pointed out that participants could follow the entire play with good understanding of both scene rationale and crowd emotion, with results pointing that these two components are crucial drivers for user overall satisfaction. Subtitles had good reviews with lesser complaints about about subtitle synchronization, and images had bad reviews mostly because of camera and hardware limitations. In general, subtitle systems for theaters are well accepted by DHH spectators and this can improve DHH people experience watching the play. This points VR and AR devices as cost reduction alternatives for accessibility in theaters and possibly other live events. All cited results were listed in the section 7.1 ("Experiments and Results related to Text Subtitling").

Based on results these finding and to enable access for DHH people with don't know to read texts, a new study detailed in the section 4.2 ("Bringing Sign Language to Live Theaters") was performed to add new improvements in proposed method, as to design a sign language subtitling and to use semantic similarity in speech correction module. The sign language subtitling, which despite neutral evaluation from performed experiments, more studies is needed to check this promising innovation. Moreover, based on collected feedback, a huge effort is needed to start a standard for displaying definition in a VR/AR device. The major pain point identified was the sign language subtitling interpreter acting was not synchronized with actors regarding time, expression and feeling demonstration. All results were detailed in the section 7.2 ("Experiments and Results related to Sign Language Subtitling").

In the chapter 5 ("Experiments and Results related to ASR") is showed the results from accuracy and performance of the modules for Automatic Speech Recognition (Google and CPqD), for all cases the Google ASR presented better results except for the real time factor which mean score represents that in average using such API would take almost the half of the spoken time to recognize a given sentence. The CPqD ASR took less time to response than the Google ASR. Given the conditions in which the method is meant to function, in which a higher delay in subtitles generation would difficult understanding given the effect on synchronization, it was selected the faster ASR. The output of the ASR is inputed to another module intended to correct such recognized text by finding a similar sentence in the script of the play. Given the score

of WRR presented by each ASR and considering that few to none improvisation is performed by an actor using the solution, it is recommended that the module for sentence correction is able to find a sentence in the script receiving as input at least 80% of the sentence's original content. However a higher flexibility can be attained dependent on the capacity of such method to recognize semantic similarity.

The evaluation of semantic similarity integration is detailed in the chapter 6 ("Experiments and Results related to Speech Correction through Semantic Similarity"), where the new speech correction module was applied to five datasets, which were submitted to two different ASRs and its outputs to the module of Speech Correction considering three groups of pairs to observe the semantic and syntactic errors presented by the modules as well as its performance under different configurations representative of possible scenarios of actual theatrical plays. The ASRs presented an error of 50% in average when applied to the audio of an actual play. The module for semantic similarity presented an average error of 18% on sentences not modified by the ASR. However, the output of the Semantic Similarity module is affected by the error introduced by the ASRs. Speech Correction based on Semantic Similarity would present less error than Syntactically based. Execution times for Semantic Similarity module did not exceed 150 milliseconds.

The current findings help us to state that VR is about to set new standards for a broad spectrum of the accessibility field as an assistive tool, unlocking useful environments for persons with disabilities while empowering them to perform regular tasks and do better through rich augmented and virtual experiences.

8.1 Future Works

With results reveling opportunity for improvements in sign language subtitles and image areas, new researches can be conducted with AR devices, where this new hardware must perform better and overcome most of current image limitations, because AR device usage will prevent any smartphones camera issues, eye strain and less concern about how stage lighting influences in camera, so DHH users will be able to see the play as it is, and the only virtual object will be the subtitle, this is a more natural and less tiring type of interaction.

Based on this thesis analysis, sign language subtitling needs to be improved, the main concern is regarding the interpreter acting, which must be synchronized with actors regarding time, expression and feeling demonstration. Then, improve the interpreter synchronization can be a interesting future work.

Another future work to be highlighted is to evaluate this method in other live scenario, where a script is followed. Moreover, the method can be upgraded to build a more complete accessibility tool that helps DHH people not only in theaters but in many other tasks of their daily lives and to improve their communication and interaction with hearing people.

Bibliography

1 CPQD. 2017. <https://speechweb.cpqd.com.br/asr/docs/2.9/get_started/prod_overview.html>. Accessed: 2019-02-10. 8, 23

2 MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. USA: Curran Associates Inc., 2013. (NIPS'13), p. 3111–3119. Disponível em: <http://dl.acm.org/citation.cfm?id=2999792.2999959>. 8, 27

3 WORD2VEC. 2019. <https://www.tensorflow.org/tutorials/representation/word2vec>. Accessed: 2019-02-10. 8, 28

4 UNIVERSALDEPENDENCIES. 2018. <https://universaldependencies.org/introduction.html>. Accessed: 2018-12-19. 8, 29

5 REARWINDOW. 2018. <https://en.wikipedia.org/wiki/Rear_Window_Captioning_System>. Accessed: 2018-12-19. 8, 38

6 DCMP. 2018. <https://dcmp.org/learn/34-x>. Accessed: 2018-12-19. 8, 39, 40

7 PROPOR. 2016. <http://propor2016.di.fc.ul.pt/?page_id=381>. Accessed: 2019-02-10. 10, 71, 72

8 ORGANIZATION, W. H. *World Report on Disability*. 2011. <http://www.who.int/disabilities/world_report/2011/report.pdf>. [Online; accessed 05-May-2016]. 16

9 CHAN, W. et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: IEEE. *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. [S.l.], 2016. p. 4960–4964. 21

10 BESACIER, L. et al. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, Elsevier, v. 56, p. 85–100, 2014. 21

11 HINTON, G. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, IEEE, v. 29, n. 6, p. 82–97, 2012. 21

12 RABINER, L. R.; JUANG, B.-H. *Fundamentals of speech recognition*. [S.l.]: PTR Prentice Hall Englewood Cliffs, 1993. v. 14. 21

13 BAKER, J. K. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, ASA, v. 65, n. S1, p. S132–S132, 1979. 22

14 VARGA, A.; STEENEKEN, H. J. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, Elsevier, v. 12, n. 3, p. 247–251, 1993. 22

15 HIRSCH, H.-G.; PEARCE, D. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*. [S.l.: s.n.], 2000. 22

16 GOOGLECLOUD. 2019. <https://cloud.google.com/speech-to-text/l>. Accessed: 2019-02-10. 22

17 ARAI, M.; KELLER, F. The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, Taylor & Francis, v. 28, n. 4, p. 525–560, 2013. 24

18 JURAFSKY, D. Speech and language processing: An introduction to natural language processing. *Computational linguistics, and speech recognition*, Prentice Hall, 2000. 24

19 GROSJEAN, F. *How long is the sentence? Prediction and prosody in the on-line processing of language*. [S.l.]: Walter de Gruyter, Berlin/New York, 1983. 24

20 WANG, J. et al. Sentence recognition from articulatory movements for silent speech interfaces. In: IEEE. *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. [S.l.], 2012. p. 4985–4988. 24

21 KIM, A. E.; OINES, L. D.; SIKOS, L. Prediction during sentence comprehension is more than a sum of lexical associations: the role of event knowledge. *Language, Cognition and Neuroscience*, Taylor & Francis, v. 31, n. 5, p. 597–601, 2016. 24

22 CUTLER, A.; FOSS, D. J. On the role of sentence stress in sentence processing. *Language and Speech*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 1–10, 1977. 25

23 GRISONI, L.; MILLER, T. M.; PULVERMÜLLER, F. Neural correlates of semantic prediction and resolution in sentence processing. *Journal of Neuroscience*, Soc Neuroscience, v. 37, n. 18, p. 4848–4858, 2017. 25

24 CORLEY, C.; MIHALCEA, R. Measuring the semantic similarity of texts. In: *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (EMSEE '05), p. 13–18. Disponível em: <http://dl.acm.org/citation.cfm?id=1631862.1631865>. 25

25 RUS, V. et al. Semilar: The semantic similarity toolkit. In: . [S.l.: s.n.], 2013. 25

26 MIHALCEA, R.; CORLEY, C.; STRAPPARAVA, C. Corpus-based and knowledge-based measures of text semantic similarity. In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*. AAAI Press, 2006. (AAAI'06), p. 775–780. ISBN 978-1-57735-281-5. Disponível em: <http://dl.acm.org/citation.cfm?id=1597538.1597662>. 25

27 CROCKETT, K. et al. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, IEEE Computer Society, Los Alamitos, CA, USA, v. 18, n. 08, p. 1138–1150, aug 2006. ISSN 1041-4347. 25

28 ISLAM, A.; INKPEN, D. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data*, ACM, New York, NY, USA, v. 2, n. 2, p. 10:1–10:25, jul. 2008. ISSN 1556-4681. Disponível em: <http://doi.acm.org/10.1145/1376815.1376819>. 26

29 MAJUMDER, G. et al. Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas*, v. 20, 2016. 26

30 RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, Nature Publishing Group, v. 323, n. 6088, p. 533, 1986. 26

31 MIKOLOV, T.; LE, Q. V.; SUTSKEVER, I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013. 26

32 MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 26

33 MIKOLOV, T.; YIH, W.-t.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2013. p. 746–751. 27

34 MARNEFFE, M.-C. D. et al. Universal stanford dependencies: A cross-linguistic typology. *Proceedings of the 9Th International Conference on Language Resources and Evaluation (LREC)*, p. 4585–4592, 01 2014. 27

35 PETROV, S.; DAS, D.; MCDONALD, R. A universal part-of-speech tagset. *Computing Research Repository - CORR*, 04 2011. 27

36 ZEMAN, D. Reusable tagset conversion using tagset drivers. In: . [S.l.: s.n.], 2008. 28

37 SILVEIRA, N. et al. A gold standard dependency corpus for English. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. [S.l.: s.n.], 2014. 28, 69

38 NOTHMAN, J. et al. Learning multilingual named entity recognition from Wikipedia. 10 2017. Disponível em: <https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500>. 28, 69

39 CHIN, H.; BERNARD-OPITZ, V. Teaching conversational skills to children with autism: Effect on the development of a theory of mind. *Journal of autism and developmental disorders*, v. 30, p. 569–83, 01 2001. 30

40 BROWN, D. et al. Development and evaluation of the virtual city. *International Journal of Virtual Reality*, Nottingham Trent University, v. 4, n. 1, p. 28–41, 1999. 30

41 BOZGEYIKLI, L. L. et al. Effects of virtual reality properties on user experience of individuals with autism. *ACM Trans. Access. Comput.*, ACM, New York, NY, USA, v. 11, n. 4, p. 22:1–22:27, nov. 2018. ISSN 1936-7228. Disponível em: <http://doi.acm.org/10.1145/3267340>. 30

42 PENNINGTON, B.; OZONOFF, S. Executive functions and development psychology. *Journal of child psychology and psychiatry, and allied disciplines*, v. 37, p. 51–87, 02 1996. 30

43 ZHANG, L. et al. Design and evaluation of a collaborative virtual environment (comove) for autism spectrum disorder intervention. *ACM Trans. Access. Comput.*, ACM, New York, NY, USA, v. 11, n. 2, p. 11:1–11:22, jun. 2018. ISSN 1936-7228. Disponível em: <http://doi.acm.org/10.1145/3209687>. 30

44 PARSONS, S.; MITCHELL, P. The potential of virtual reality in social skills training for people with autistic spectrum disorders. *Journal of intellectual disability research*, Wiley Online Library, v. 46, n. 5, p. 430–443, 2002. 30

45 KRANTZ, P. J.; MCCLANNAHAN, L. E. Social interaction skills for children with autism: A script-fading procedure for beginning readers. *Journal of applied behavior analysis*, v. 31, p. 191–202, 02 1998. 30

46 ZHAO, H. et al. Design of a haptic-gripper virtual reality system (hg) for analyzing fine motor behaviors in children with autism. *ACM Trans. Access. Comput.*, ACM, New York, NY, USA, v. 11, n. 4, p. 19:1–19:21, nov. 2018. ISSN 1936-7228. Disponível em: <http://doi.acm.org/10.1145/3231938>. 30

47 SVEISTRUP, H. Motor rehabilitation using virtual reality. *Journal of neuroengineering and rehabilitation*, BioMed Central, v. 1, n. 1, p. 10, 2004. 31

48 LANGE, B. et al. Designing informed game-based rehabilitation tasks leveraging advanced n virtual reality. *Disability and rehabilitation*, v. 34, p. 1863–70, 04 2012. 31

49 ELOR, A.; TEODORESCU, M.; KURNIAWAN, S. Project star catcher: A novel immersive virtual reality experience for upper limb rehabilitation. *ACM Trans. Access. Comput.*, ACM, New York, NY, USA, v. 11, n. 4, p. 20:1–20:25, nov. 2018. ISSN 1936-7228. Disponível em: <http://doi.acm.org/10.1145/3265755>. 31

50 KIM, N.; YOO, C.; IM, J. A new rehabilitation training system for postural balance control using virtual reality technology. IEEE, v. 7, p. 482–485, 1999. 31

51 SVEISTRUP, H. et al. Experimental studies of virtual reality-delivered compared to conventional exercise programs for rehabilitation. v. 6, p. 243–249, 2003. 31

52 SVEISTRUP, H. et al. Outcomes of intervention programs using flatscreen virtual reality. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, v. 7, p. 4856–8, 02 2004. 31

53 MCCOMAS, J.; SVEISTRUP, H. Virtual reality applications for prevention, disability awareness, and physical therapy rehabilitation in neurology: our recent work. v. 26, p. 55–61, 2002. 31

54 HOWE, J. et al. The community balance and mobility scale - a balance measure for individuals with traumatic brain injury. *Clinical rehabilitation*, v. 20, p. 885–95, 11 2006. 31

55 G, R. Virtual reality in paraplegia: a vr-enhanced orthopedic appliance for walking and rehabilitation. Amsterdam, p. 209–218, 1998. 31

56 KIZONY, R.; KATZ, N. et al. Adapting an immersive virtual reality system for rehabilitation. *Computer Animation and Virtual Worlds*, Wiley Online Library, v. 14, n. 5, p. 261–268, 2003. 31

57 ARAN SEDEF SAHIN, B. T. T. D. O. T.; KAYISHAN, H. Virtual reality and occupational therapy. In: ____. [S.l.: s.n.], 2017. ISBN 978-953-51-3321-6. 31

58 MIRELMAN, A. et al. Virtual reality for gait training: can it induce motor learning to enhance complex walking and reduce fall risk in patients with parkinson's disease? *The Journals of Gerontology: Series A*, Oxford University Press, v. 66, n. 2, p. 234–240, 2011. 31

59 MIRZAEI, M. R.; GHORSHI, S.; MORTAZAVI, M. Combining augmented reality and speech technologies to help deaf and hard of hearing people. In: IEEE. *Virtual and Augmented Reality (SVR), 2012 14th Symposium on*. [S.l.], 2012. p. 174–181. 32, 34

60 BERKE, L. Displaying confidence from imperfect automatic speech recognition for captioning. *ACM SIGACCESS Accessibility and Computing*, ACM, n. 117, p. 14–18, 2017. 32, 34

61 PIQUARD-KIPFFER, A. et al. Qualitative investigation of the display of speech recognition results for communication with deaf people. In: *6th Workshop on Speech and Language Processing for Assistive Technologies*. [S.l.: s.n.], 2015. p. 7. 32, 34

62 LUO, X. et al. Assistive learning for hearing impaired college students using mixed reality: a pilot study. In: IEEE. *Virtual Reality and Visualization (ICVRV), 2012 International Conference on*. [S.l.], 2012. p. 74–81. 32, 35

63 KERCHER, K.; ROWE, D. C. Improving the learning experience for the deaf through augment reality innovations. In: IEEE. *Engineering, Technology and Innovation (ICE), 2012 18th International ICE Conference on*. [S.l.], 2012. p. 1–11. 32, 36

64 SILVA, Y. M. et al. Training environment for electric powered wheelchairs using teleoperation through a head mounted display. In: . [S.l.: s.n.], 2018. p. 1–2. 32, 33

65 TEóFILO, M. et al. Bringing basic accessibility features to virtual reality context. In: *2016 IEEE Virtual Reality (VR)*. [S.l.: s.n.], 2016. p. 293–294. ISSN 2375-5334. 32

66 Teófilo, M. et al. Evaluating accessibility features designed for virtual reality context. In: *2018 IEEE International Conference on Consumer Electronics (ICCE) (2018 ICCE)*. Las Vegas, USA: [s.n.], 2018. 32, 33

67 TEÓFILO, M. et al. Exploring virtual reality to enable deaf or hard of hearing accessibility in live theaters: A case study. In: ANTONA, M.; STEPHANIDIS, C. (Ed.). *Universal Access in Human-Computer Interaction. Virtual, Augmented, and Intelligent Environments*. Cham: Springer International Publishing, 2018. p. 132–148. ISBN 978-3-319-92052-8. 32, 45

68 GUO, R.; SAMARAWEERA, G.; QUARLES, J. A unique way to increase presence of mobility impaired users 2014; increasing confidence in balance. In: *Virtual Reality (VR), 2014 iEEE*. [S.l.: s.n.], 2014. p. 77–78. 32

69 GUO, R.; SAMARAWEERA, G.; QUARLES, J. Mobility impaired users respond differently than healthy users in virtual environments. *Computer Animation and Virtual Worlds*, p. n/a–n/a, 2014. ISSN 1546-427X. Disponível em: <http://dx.doi.org/10.1002/cav.1610>. 32

70 GUO, R.; SAMARAWEERA, G.; QUARLES, J. The effects of avatars on presence in virtual environments for persons with mobility impairments. In: NOJIMA, T.; REINERS, D.; STAADT, O. (Ed.). *ICAT-EGVE 2014 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. [S.l.]: The Eurographics Association, 2014. ISBN 978-3-905674-65-1. ISSN 1727-530X. 32

71 GUO, R.; SAMARAWEERA, G.; QUARLES, J. The effects of ves on mobility impaired users: Presence, gait, and physiological response. In: *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*. New York, NY, USA: ACM, 2013. (VRST '13), p. 59–68. ISBN 978-1-4503-2379-6. Disponível em: <http://doi.acm.org/10.1145/2503713.2503719>. 32

72 SAMARAWEERA, G.; GUO, R.; QUARLES, J. Latency and avatars in virtual environments and the effects on gait for persons with mobility impairments. In: *3D User Interfaces (3DUI), 2013 IEEE Symposium on*. [S.l.: s.n.], 2013. p. 23–30. 32

73 SAMARAWEERA, G.; GUO, R.; QUARLES, J. Head tracking latency in virtual environments revisited: Do users with multiple sclerosis notice latency less? IEEE, 2015. 33

74 RODRIGUEZ, N. Development of a wheelchair simulator for children with multiple disabilities. In: *Virtual and Augmented Assistive Technology (VAAT), 2015 3rd IEEE VR International Workshop on*. [S.l.: s.n.], 2015. p. 19–21. 33

75 CANTU, M. et al. Game cane: An assistive 3dui for rehabilitation games. In: *3D User Interfaces (3DUI), 2014 IEEE Symposium on*. [S.l.: s.n.], 2014. p. 43–46. 33

76 LV, Z. et al. A game based assistive tool for rehabilitation of dysphonic patients. *CoRR*, abs/1504.01030, 2015. Disponível em: <http://arxiv.org/abs/1504.01030>. 33

77 COSTA, R. T. d.; CARVALHO, M. R. d.; NARDI, A. E. Virtual reality exposure therapy in the treatment of driving phobia. *Psicologia: Teoria e Pesquisa*, SciELO Brasil, v. 26, n. 1, p. 131–137, 2010. 33

78 HONG, R. et al. Dynamic captioning: video accessibility enhancement for hearing impairment. In: ACM. *Proceedings of the 18th ACM international conference on Multimedia*. [S.l.], 2010. p. 421–430. 35

79 BEADLES, R. L.; BALL, J. E. *Method and apparatus for closed captioning at a performance*. [S.l.]: Google Patents, 1997. US Patent 5,648,789. 35

80 KIPP, M.; HALOIR, A.; NGUYEN, Q. Sign language avatars: Animation and comprehensibility. In: . [S.l.: s.n.], 2011. p. 113–126. 36

81 ADAMO-VILLANI, N.; BENI, G. Sign language subtitling by highly comprehensible "semantroids". *Journal of Educational Technology Systems*, v. 35, p. 61–87, 12 2006. 36

82 SCASSELLATI, B. et al. Teaching language to deaf infants with a robot and a virtual human. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2018. (CHI '18), p. 553:1–553:13. ISBN 978-1-4503-5620-6. Disponível em: <http://doi.acm.org/10.1145/3173574.3174127>. 36

83 MONTAGUD, M. et al. Imac: Enabling immersive, accessible and personalized media experiences. In: . [S.l.: s.n.], 2018. p. 245–250. 36

84 VINAYAGAMOORTHY, V. et al. Personalising the tv experience with augmented reality technology: Synchronised sign language interpretation. In: *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video*. New York, NY, USA: ACM, 2018. (TVX '18), p. 179–184. ISBN 978-1-4503-5115-7. Disponível em: <http://doi.acm.org/10.1145/3210825.3213562>. 37, 94

85 UCHIDA, T. et al. Sign language support system for viewing sports programs. In: . [S.l.: s.n.], 2017. p. 339–340. 37

86 UCHIDA, T. et al. Evaluation of a sign language support system for viewing sports programs. In: *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, NY, USA: ACM, 2018. (ASSETS '18), p. 361–363. ISBN 978-1-4503-5650-3. Disponível em: <http://doi.acm.org/10.1145/3234695.3241002>. 37

87 MARTINO, J. a. D. et al. Building a brazilian portuguese - brazilian sign language parallel corpus using motion capture data. In: . [S.l.: s.n.], 2016. 37

88 BUTLER, J. Perspectives of deaf and hard of hearing viewers of captions. *American Annals of the Deaf*, v. 163, p. 534–553, 01 2019. 37

89 ROTHE, S.; TRAN, K.; HUSSMANN, H. Dynamic subtitles in cinematic virtual reality. In: *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video*. New York, NY, USA: ACM, 2018. (TVX '18), p. 209–214. ISBN 978-1-4503-5115-7. Disponível em: <http://doi.acm.org/10.1145/3210825.3213556>. 41

90 GUIMARAES, R. L.; BRITO, J. O.; SANTOS, C. A. S. Investigating the influence of subtitles synchronization in the viewer's quality of experience. In: *Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2018. (IHC 2018), p. 30:1–30:10. ISBN 978-1-4503-6601-4. Disponível em: <http://doi.acm.org/10.1145/3274192.3274222>. 41

91 KAFLE, S.; HUENERFAUTH, M. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In: *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, NY, USA: ACM, 2017. (ASSETS '17), p. 165–174. ISBN 978-1-4503-4926-0. Disponível em: <http://doi.acm.org/10.1145/3132525.3132542>. 42

92 CENSO. 2010. <https://biblioteca.ibge.gov.br/visualizacao/periodicos/94/cd_2010_religiao_deficiencia.pdf>. Accessed: 2019-02-26. 42

93 NATIONAL. 2019. <https://www.nationaltheatre.org.uk/your-visit/access/caption-glasses>. Accessed: 2019-02-26. 41

94 BROADWAY. 2018. <https://www.broadway.com/buzz/193583/national-theatres-innovative-closed-caption-glasses-could-transform-access-to-the-arts/>. Accessed: 2019-02-26. 41

95 DOMINGUES, L. de A.; TINTO-PB, R. Cinelibras: Uma proposta para geração automática e distribuição de janelas de libras em salas de cinema. 2013. 55

96 GOOGLEVR. 2018. <https://developers.google.com/vr/design/sticker-sheet>. Accessed: 2018-12-19. 55

97 J.ARORA, S.; SINGH, R. Automatic speech recognition: A review. *International Journal of Computer Applications*, v. 60, p. 34–44, 12 2012. 60

98 HUANG, X.; ACERO, A.; HON, H.-W. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. 1st. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001. ISBN 0130226165. 60

99 NETO, N. et al. Free tools and resources for brazilian portuguese speech recognition. *Journal of the Brazilian Computer Society*, v. 17, n. 1, p. 53–68, Mar 2011. ISSN 1678-4804. Disponível em: <https://doi.org/10.1007/s13173-010-0023-1>. 60, 61, 73, 77

100 FONSECA, E.; CRISCUOLO, M.; ALUISIO, S. Assin: Avaliacao de similaridade semantica e inferencia textual. In: . [S.l.: s.n.]. 70

101 TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. *CoRR*, abs/1003.1141, 2010. Disponível em: <http://arxiv.org/abs/1003.1141>. 70

102 MARELLI, M. et al. A sick cure for the evaluation of compositional distributional semantic models. In: . [S.l.: s.n.]. 71

103 GEEN, R. Arrabal's "the architect and emperor of assyria". *Romance Notes*, University of North Carolina at Chapel Hill for its Department of Romance Studies, v. 19, n. 2, p. 140–145, 1978. ISSN 00357995, 21657599. Disponível em: <http://www.jstor.org/stable/43801555>. 89

104 M., M. D. D.; KOZUH, I. A comparison of processes in sign language interpreter videos with or without captions. *PLOS ONE*, Public Library of Science, v. 10, n. 5, p. 1–15, 05 2015. Disponível em: <https://doi.org/10.1371/journal.pone.0127577>. 95