FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Context Analysis for Voicing Decision in Whispered Speech

Gonçalo Amaral Tavares

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Aníbal João de Sousa Ferreira

October 10, 2024

Abstract

Whispered Speech (WS) is the result of an absence of vibrations in the vocal folds during speech. While it has its uses in modern-day communication, certain individuals are forced to use it due to external factors such as speech impediments, which can lead to low self-esteem and poor mental stability. To combat this, several projects were created with the goal of converting WS to normal speech. This dissertation explores the challenges and solutions associated with processing WS for improved voice recognition systems. The primary objective of this research is to enhance the accuracy of voicing decision (VD) in WS using context-based machine learning and deep learning methodologies. Methodologically, the research employs a comprehensive approach that includes data acquisition, preprocessing, feature engineering, and model training. Specifically, a dataset was created using audio samples from multiple speakers, and features such as Mel-frequency Cepstral Coefficients (MFCCs) were extracted and normalized. A series of machine learning models, including Logistic Regression (LR), Random Forest (RF), and various neural network architectures, were developed and evaluated. The performance of these models was assessed based on their accuracy and precision in identifying candidate segments to voicing within the WS dataset. The study further investigates the impact of context-based analysis, testing models with and without contextual information to determine the optimal conditions for VD. The results indicate significant improvements in model performance when incorporating contextual information. Without context, the RF model achieved the highest accuracy at 85.79%, with a precision of 83.85%, recall of 88.71%, F1 score of 86.21%, and AUC of 93.77%. When context was included, the Long Short-Term Memory (LSTM) model demonstrated the best performance with an accuracy of 95.93%, precision of 95.92%, recall of 95.94%, F1 score of 95.93%, and AUC of 99.31%. These findings underscore the importance of context in enhancing the accuracy of VD in WS. The research successfully demonstrates that advanced machine learning techniques, coupled with context-based analysis, can significantly improve the processing and recognition of whispered speech. Future work will focus on refining these models and dataset.

Key Words: whispered speech, voicing decision, context analysis, voiced, unvoiced

Resumo

O discurso sussurrado é o resultado da ausência de vibrações nas cordas vocais durante a fala. Embora tenha os seus usos na comunicação em tempos modernos, certos indivíduos são forçados a utilizá-lo devido a fatores externos, como distúrbios da fala, o que pode levar a uma baixa autoestima e instabilidade mental. A procura de uma solução levou à criação de vários projetos com o objetivo de converter o discurso sussurrado em fala normal. Esta dissertação explora os desafios e soluções associados ao processamento do discurso sussurrado para melhorar os sistemas de reconhecimento de voz. O objetivo principal desta pesquisa é aprimorar a precisão da decisão de vozeamento no discurso sussurrado utilizando metodologias de machine learning e deep learning baseadas em contexto. Metodologicamente, a pesquisa adota uma abordagem abrangente que inclui aquisição de dados, pré-processamento dos mesmos, feature engineering e treino de modelos. Especificamente, foi criado um conjunto de dados utilizando amostras de áudio de múltiplos oradores, e características como os Mel-frequency Cepstral Coefficients foram extraídas e normalizadas. Uma série de modelos de aprendizagem automática, incluindo Logistic Regression, Random Forest e várias arquiteturas de redes neurais, foram desenvolvidos e avaliados. O desempenho destes modelos foi avaliado com base na sua precisão e exatidão na identificação de segmentos candidatos a vozeamento dentro do conjunto de dados de discurso sussurrado. O estudo investiga ainda o impacto da análise baseada em contexto, testando modelos com e sem informação contextual para determinar as condições óptimas para a decisão de vozeamento. Os resultados indicam melhorias significativas no desempenho dos modelos ao incorporar informação contextual. Sem contexto, o modelo de Random Forest alcançou a maior precisão de 85.79%, com uma exatidão de 83.85%, recall de 88.71%, F1 score de 86.21%, e AUC de 93.77%. Quando o contexto foi incluído, o modelo de Long Short-Term Memory demonstrou o melhor desempenho, com uma precisão de 95.93%, exatidão de 95.92%, recall de 95.94%, F1 score de 95.93%, e AUC de 99.31%. Estes resultados comprovam a importância do contexto na melhoria da precisão da decisão de vozeamento no discurso sussurrado. A pesquisa demonstra com sucesso que técnicas avançadas de machine learning, aliadas a uma análise baseada em contexto, podem melhorar significativamente o processamento e o reconhecimento do discurso sussurrado. Trabalhos futuros deverão focar-se no aperfeiçoamento destes modelos e do *dataset*.

Palavras Chave: discurso sussurado, decisão de vozeamento, análise contextual, vozeado, não-vozeado

SDG in Master's Dissertation

The dissertation's potential impacts on the United Nations' Sustainable Development Goals (SDGs) are described in more detail in Appendix A. These include contributions to health technology under SDG 3, promotion of equal education access under SDG 4, innovation in industry under SDG 9, reduction of inequalities under SDG 10, and fostering global partnerships under SDG 17. Key effort indicators include research and development investments, prototype testing, user feedback, academic dissemination, training programs, collaborations, and policy advocacy.

Agradecimentos

Ao Professor Aníbal João de Sousa Ferreira, o meu orientador, quero agradecer o tempo e assistência que me deu quando e sempre que necessitei.

Ao Engenheiro João Miguel Pinto Pereira da Silva, o meu supervisor, quero agradecer a paciência, dedicação e disponibilidade que teve comigo no decorrer deste trabalho.

À Paula, minha mãe, e ao Mário, meu pai, quero agradecer todo o encorajamento oferecido e sacrifícios feitos nestes 5 anos. Sem eles, ter-me-ia com certeza desviado do caminho em alguma altura.

À minha família, quero agradecer pelo apoio e confiança que me manteve forte durante tempos difíceis.

Aos meus amigos, quero agradecer a própria presença na minha vida, os laços que nós forjámos sempre serão um dos meus bens mais valiosos.

Aos meus avôs e ao meu tio, quero agradecer por me terem acompanhado durante a esta jornada, e prometo manter-vos sempre comigo depois de chegar à sua conclusão.

Gonçalo Amaral Tavares

“All human wisdom is contained in these two words – "Wait and Hope".”

Alexandre Dumas in *"The Count of Monte Cristo"*

Contents

1 Introduction **1**
- 1.1 Contextualization and Motivation . . . 1
- 1.2 Problem to Solve . . . 1
- 1.3 Dissertation Goals . . . 2
- 1.4 Methodology Used . . . 2
- 1.5 The Outcomes . . . 2
- 1.6 Outline of the Report . . . 2
- 1.7 Chapter Summary . . . 4

2 Background **5**
- 2.1 Speech Processing: Basic Concepts . . . 5
  - 2.1.1 Fourier transform . . . 5
  - 2.1.2 Mel-frequency cepstral coefficients . . . 7
  - 2.1.3 Fundamental frequency . . . 7
  - 2.1.4 Linear predictive coding . . . 7
- 2.2 Normal Speech . . . 8
- 2.3 Whispered Speech . . . 8
- 2.4 Artificial Intelligence . . . 8
  - 2.4.1 Machine learning algorithms . . . 9
  - 2.4.2 Deep learning algorithms . . . 10
- 2.5 Chapter Summary . . . 11

3 Related Works **12**
- 3.1 Paper Selection . . . 12
- 3.2 Review of Selected Papers . . . 12
  - 3.2.1 Rule-based approach using spectral centroid thresholding . . . 12
  - 3.2.2 Rule-based approach combining context analysis to different types of thresholding . . . 13
  - 3.2.3 LSTM-based approach applied to DNN . . . 13
  - 3.2.4 DNN-based approach, using MFCCs as training parameters . . . 14
  - 3.2.5 GMM-based approach to NAM-to-speech conversion . . . 14
  - 3.2.6 Hybrid approach using a KNN phoneme classifier followed by rule-based voicing decision using spectral centroid thresholding . . . 15
  - 3.2.7 DL-based approaches to WS segmentation . . . 15
- 3.3 Chapter Summary . . . 16

4 Methodology **17**
4.1 Hardware and Software . . . 17
4.1.1 Hardware . . . 17
4.1.2 Software . . . 17
4.2 Specifics on the Dataset . . . 18
4.2.1 Selection process and recording environment . . . 18
4.2.2 Dataset structure . . . 18
4.3 Dataset Preprocessing . . . 18
4.3.1 Downsampling . . . 18
4.3.2 Phone-based annotation and segmentation . . . 18
4.3.3 Selection and cleaning . . . 19
4.3.4 Labelling segments as voicing candidates . . . 19
4.3.5 Normalization of audio segments . . . 20
4.4 Feature Engineering . . . 21
4.4.1 Feature extraction . . . 21
4.4.2 Feature normalization . . . 21
4.4.3 Dataset explosion from segments to frames . . . 21
4.4.4 Class rebalancing through selective silence frame reduction . . . 21
4.4.5 Context-based analysis . . . 22
4.4.6 Baseline definition . . . 22
4.5 Evaluation Metrics Definition . . . 22
4.5.1 Accuracy . . . 22
4.5.2 Precision . . . 22
4.5.3 Recall . . . 23
4.5.4 F1 Score . . . 23
4.5.5 Area Under the ROC Curve . . . 23
4.6 Assessment of the Best Model . . . 23
4.7 Chapter Summary . . . 23

5 Results and Discussion **24**
5.1 Dataset Aquisition . . . 24
5.2 Dataset Preprocessing (Results) . . . 24
5.3 Feature Engineering (Results) . . . 25
5.3.1 Feature extraction . . . 25
5.3.2 Feature normalization . . . 25
5.3.3 Dataset explosion from segments to frames . . . 25
5.3.4 Class rebalancing through selective silence frame reduction . . . 25
5.3.5 Context-based analysis . . . 26
5.3.6 Baseline definition . . . 26
5.4 Assessment of the Best Model (Results) . . . 26
5.4.1 Results without context . . . 26
5.4.2 Results with context . . . 28
5.4.3 Analysis of context intervals . . . 30
5.5 Chapter Summary . . . 31

6 Conclusions and Future Work **32**
6.1 Summary of Conclusions 32
6.1.1 Data preprocessing 32
6.1.2 Feature engineering 32
6.1.3 Assessment of the best performing model 32
6.1.4 Analysis of context intervals 33
6.2 Fulfillment of the Intended Goals 33
6.3 Contributions and Limitations 33
6.3.1 Future work 34
6.4 Chapter Summary 34

References **35**

A SDG in Master's Dissertation **37**

List of Figures

2.1 Graphs displaying illustrative time signals and Fourier Transforms. 6
(a) Visual Representation of the FT [3] 6
(b) Example of the DFT [3] 6
2.2 Example of a multilayer perceptron 10

5.1 Classifiers' performance metrics obtained from 10 iterations of TTS evaluation without context. 28
5.2 Classifiers' performance metrics obtained from 10 iterations of TTS evaluation with context. 30
5.3 Classifiers' performance metrics obtained from 10 iterations of TTS evaluation with sliding context window. 31

List of Tables

4.1 SAMPA phonetic annotation labels of segments and candidate to voicing label. . 20

5.1 Classifiers' performance metrics and standard deviations obtained from 10 iterations of TTS evaluation without context. 28
5.2 Classifiers' performance metrics and standard deviations obtained from 10 iterations of TTS evaluation with context. 30

Abreviaturas e Símbolos

List of abbreviations

AI	Artificial Intelligence
AUC	Area Under the ROC Curve
CAPE-V	Consensus Auditory-Perceptual Evaluation of Voice
CFT	Continuous Fourier Transform
CMU	Carnegie Mellon University
CNN	Convolutional Neural Network
CTV	Candidate to Voicing
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DL	Deep Learning
DNN	Deep Neural Network
EL	Ensemble Learning
EP	European Portuguese
FFT	Fast Fourier Transform
FS	Frame Size
FT	Fourier Transform
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Unit
HS	Hop Size
ID	Identification
KNN	K-Nearest Neighbors
LPC	Linear Predictive Coding
LR	Logistic Regression
LSTM	Long Short-Term Memory
MIR-1K	Multimedia Information Retrieval lab, 1000 song clips
MFCCs	Mel-frequency Cepstral Coefficients
ML	Machine Learning
MLP	Multilayer Perceptron
NAM	Non Audible Murmur
NCTV	Not Candidate to Voicing
NS	Normal Speech
PAL	Phonetic Annotation Label

RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SAMPA	Speech Assessment Methods Phonetic Alphabet
SDG	Sustainable Development Goal
SM	Speech Mode
SPSS	Statistical Parametric Speech Synthesis
SWAV	Segment's Waveform Audio File Format File
TCN	Temporal Convolutional Network
TTS	Train-Test Split
UN	United Nations
VD	Voicing Decision
WAV	Waveform Audio File Format
WS	Whispered Speech

List of symbols

f_0	Fundamental Frequency

Chapter 1

Introduction

This chapter outlines the context and motivation behind the project, defining the overall objectives of the dissertation. It briefly goes over the methodology used, gives a summary of the conclusions that were reached and concludes with a brief description of the report's structure.

1.1 Contextualization and Motivation

Whispered speech (WS), commonly known as whispering, is the act of talking quietly, which results from an absence of vibrations in the vocal folds during speech production. The use of WS occurs in modern-day communication, specifically in situations where privacy and secrecy are expected or even mandatory. However, certain external factors, such as speech impediments that affect the vocal folds, can hinder individuals, forcing them to communicate using exclusively WS. This often leads to a loss of intelligibility, a weak presence in conversations, and consequently, poor mental stability and low self-esteem. To address the aforementioned issues, researchers have embarked on the development of technology capable of converting WS into normal speech (NS), which in turn led to the rise of several different projects and approaches to the subject such as *"DyNaVoiceR"*, a project focused towards converting whispers to NS in order to combat aphonia.

1.2 Problem to Solve

Speech conversion techniques tend to vary on approach, but their purpose is fundamentally the same, to provide the missing vibrational components WS lacks and reconstruct it as NS. However, not every phone in WS requires said components, which complicates the process, as it must abide by a careful selection of which phones to convert known as voicing decision (VD). VD is typically realized on a context-free basis, that is, without regard of what segments are considered voicing candidates or not. In spite of this, a previous study showed that the use of context in the decision making could improve the overall results by a large margin [15].

1.3 Dissertation Goals

The goal of this dissertation is to contribute findings on the importance a context analysis can have on the VD process, and document the effect it has on different methods of VD. The specific objectives are listed as follows:

1. **Description of a phonetically annotated speech dataset:** Describe the acquisition and characteristics of the available phonetically annotated dataset;
2. **Preprocessing of the Dataset:** Prepare the dataset for feature engineering;
3. **Feature Engineering:** Extract features from the preprocessed dataset, Process them to be amenable for subsequent analysis and define a baseline;
4. **Evaluation metrics definition:** Define performance metrics to quantitatively evaluate the models;
5. **Assessment of the best performing model:** Judge and compare the models based on the established criteria and presence or lack of context, and determine which one is the best;
6. **Analysis of context intervals:** Further analyse the best performing model and determine at which context interval it performs best.

1.4 Methodology Used

For this study various machine learning (ML) and deep learning (DL) models made predictions on a preprocessed dataset consisting of audio features in which the one with the best overall accuracy and precision would be considered the most effective.

1.5 The Outcomes

The dissertation was able to simulate a VD algorithm in a situation without context and one with context. The context-based environment had a better score all across the board, the long-short term memory (LSTM) model in particular being assessed as the best model for a context approach to VD.

1.6 Outline of the Report

This dissertation report is structured as follows:

1 **Introduction:** This chapter introduces the report, featuring chapters on its:

1.1 **Contextualization and Motivation;**

1.2 **Problem to Solve;**

1.3 **Dissertation Goals;**

1.4 **Methodology Used;**

1.5 **The Outcomes.**

2 **Background:** This chapter focuses on the scientific know-how required to understand the project. It includes sections on:

2.1 **Speech Processing: Basic Concepts:** Describes basic knowledge for speech processing such as the fourier transform 2.1.1, mel-frequency cepstral coefficients 2.1.2, fundamental frequency 2.1.3 and linear predictive coding 2.1.4;

2.2 **Normal Speech:** It gives in-depth information about NS;

2.3 **Whispered Speech:** Describes how WS operates;

2.4 **Artificial Intelligence:** Describes the artificial intelligence algorithms used to make predictions for this study, ML 2.4.1 and DL 2.4.2, and also describes the specific models that will be evaluated like logistic regression 2.4.1.1, random forest 2.4.1.2, multilayer perceptrons 2.4.1.3, convolutional neural network 2.4.2.1 and the two types of recurrent neural networks 2.4.2.2;

3 **Related Works:** This chapter presents a review of other papers relevant to the topic of this one. It includes sections on:

3.1 **Paper Selection:** Describes the selection process for the works;

3.2 **Review of Selected Papers:** Gives detailed explanations on how they approached VD.

4 **Methodology:** Presents in more detail the methodology used in the study, includes sections on:

4.1 **Hardware and Software:** Gives a description on the hardware 4.1.1 and software 4.1.2 employed in the study;

4.2 **Specifics of the Dataset:** Explains how the dataset was created mentioning how the speakers were selected 4.2.1 and how the dataset is structured 4.2.2;

4.3 **Dataset Preprocessing:** Describes the preprocessing steps done to the dataset including a downsampling of the audio files 4.3.1, segmentation based on phone annotation 4.3.2, selection and cleaning 4.3.3, segment labelling 4.3.4 and segment normalizing 4.3.5;

4.4 **Feature Engineering:** Gives context on the feature engineering process including feature extraction 4.4.1 and normalization 4.4.2, dataset explosion from segments to frames 4.4.3, class rebalancing via selective silence frame reduction 4.4.4, details on the context based analysis 4.4.5 and defines the baseline 4.4.6;

4.5 **Evaluation Metrics Definition:** Goes over the main evaluation metrics for the models including accuracy 4.5.1, precision 4.5.2, F1 score 4.5.4, Recall 4.5.3 and Area Under the ROC curve (ROC standing for Receiver Operating Characteristic) 4.5.5;

4.6 **Assessment of the best model:** Gives insight to the intended method for assessing which model can be considered the best.

5 **Results and Discussion:** Gives a detailed record of the results achieved, includes sections on:

5.1 **Dataset Acquisition:** Puts into practice the methodology of dataset acquisition from the previous chapter (4.2);

5.2 **Dataset Preprocessing:** Puts into practice the methodology of dataset preprocessing from the previous chapter (4.3);

5.3 **Feature Engineering:** Puts into practice the methodology of feature engineering from the previous chapter (4.4);

5.4 **Assessment of the Best Model:** Puts into practice the methodology of assessing the best model from the previous chapter (4.6);

6 **Conclusions and Future Work:** Concludes the document, including sections on:

6.1 **Summary of Conclusions:** Summarizes the contents of sections (5.1), (5.2), (5.3) and (5.4);

6.2 **Fulfillment of the intended goals**: Restates the dissertation goals from (1.3) and verifies if any were left unaccomplished;

6.3 **Contributions and limitations:** Describes what the study contributed to the area of study as well as the limitations it faced and what related future works should do to improve them (6.3.1).

1.7 Chapter Summary

In this chapter, the context and motivation of the project were outlined. The objectives were defined, making notable mention of the intended methods to carry them out as well as a preview of the resulting outcomes. It also briefly described the document structure.

The following chapter will focus on the scientific know-how required to understand the project.

Chapter 2

Background

This chapter focuses on the scientific know-how required to understand the project, tackling topics such as basic concepts for speech processing, the different types of speech that will be analysed, and the artificial intelligence (AI) algorithms used to achieve the results.

2.1 Speech Processing: Basic Concepts

Speech is inherently an analog signal. To efficiently process the audio tracks in this dissertation, it is crucial to sample and quantize those signals. These steps are essential in converting the analog signals to digital, which are more suitable for processing.

2.1.1 Fourier transform

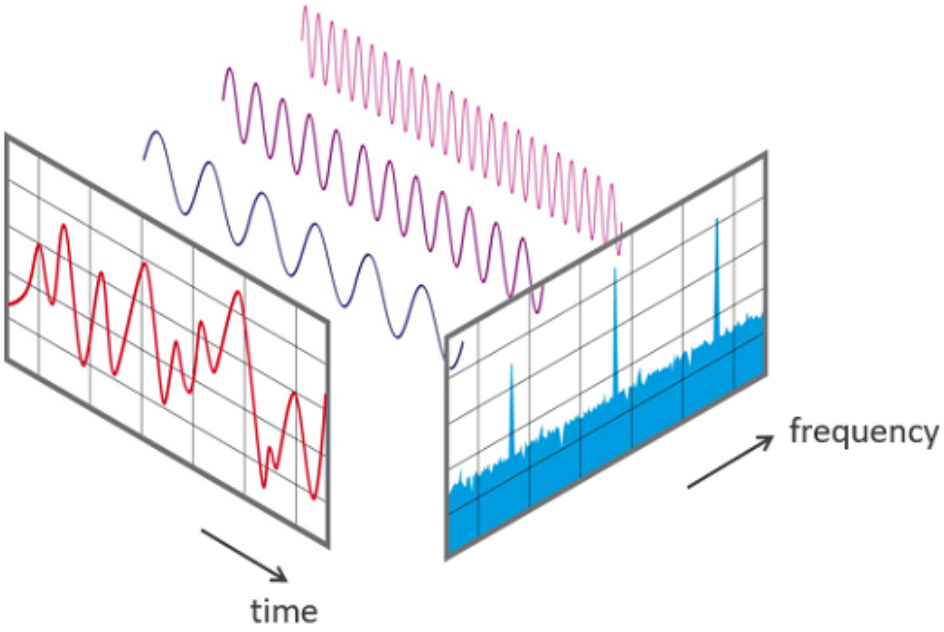
The fourier transform (FT) is widely recognized as the most popular technique for speech processing. It conceptualizes signals as linear combinations of different sinusoids, which simplifies the process of identifying frequency components of the signal to determining the frequencies, phases, and amplitudes of each sinusoid, as shown in Figure 2.1a.

The formula for the FT varies based on whether the input signal is continuous (CFT) or discrete (DFT). Assuming a continuous-time energy signal *x(t)*, and a discrete-time energy signal *x[n]*, the formulas are as in Eq. (2.1) and Eq. (2.2), respectively:

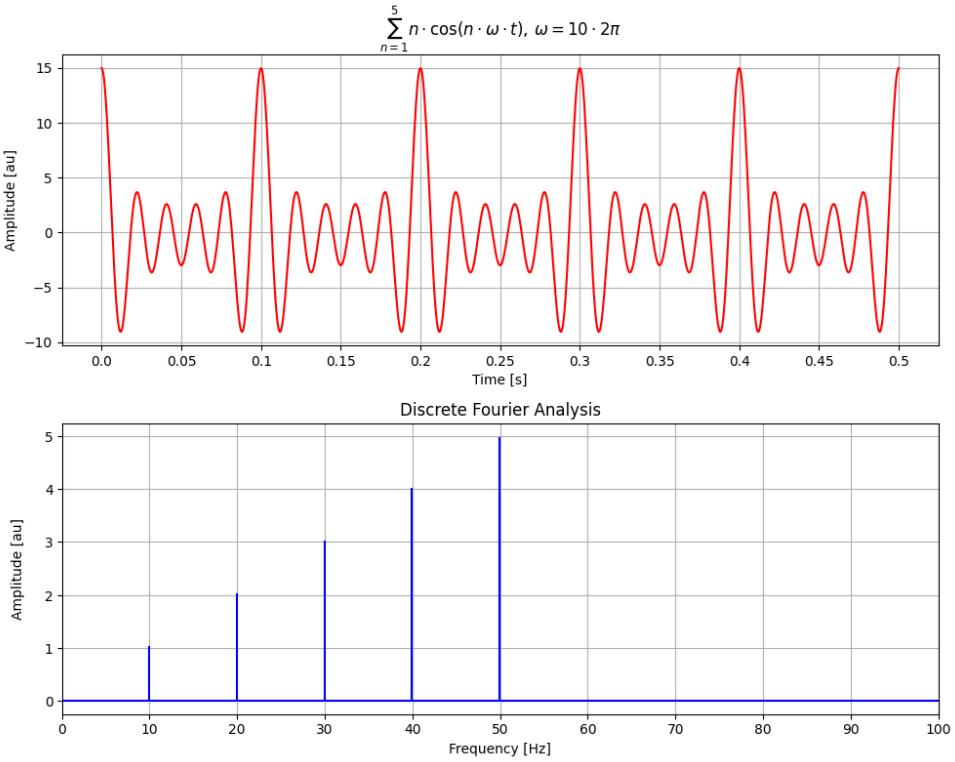
$$CFT: \quad X(f) = \int_{t=-\infty}^{+\infty} x(t) e^{-j2\pi ft}\, dt \tag{2.1}$$

$$DFT: \quad X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi nk}{N}} \tag{2.2}$$

In the case of the DFT (2.2), the k value denotes frequency index, or bin, and N denotes the total number of samples in a vector. An example is shown in Figure 2.1b.



(a) Visual Representation of the FT [3]



(b) Example of the DFT [3]

Figure 2.1: Graphs displaying illustrative time signals and Fourier Transforms.

2.1.2 Mel-frequency cepstral coefficients

Referred to as MFCCs for short, mel-frequency cepstral coefficients can be described as a set of features of a signal that provide a compact representation for the spectral envelope of the magnitude spectrum, and are often used to analyze the timbre of speech sounds.

The process to calculate MFCCs involves several steps:

1. Convert the speech signal to a digital one (if it is not already digital);
2. Pre-emphasize the signal to increase energy at high frequencies;
3. Smooth the resulting frame's edges by applying a window function (usually a Hamming Window);
4. Apply the DFT to the output signal;
5. Use the mel filterbank to decompose the signal into separate frequency bands based on the Mel scale;
6. Take the logarithm of the result;
7. Apply the discrete cosine transform (DCT) (similar to the DFT but utilizing cosine functions), to decorrelate the coefficients.

The visual representation of this process can be summarized as follows:

$DigitalSignal \rightarrow Pre-emphasize \rightarrow HammingWindow \rightarrow DFT \rightarrow MelFilterbank \rightarrow log() \rightarrow DCT$

This series of steps results in the extraction of relevant MFCCs, with higher-frequency coefficients often disregarded due to containing less relevant information.

2.1.3 Fundamental frequency

An essential element of speech processing is the fundamental frequency (represented as f_0), also known as the lowest harmonic component in voiced sound. It plays a crucial role in conveying details about the naturalness of the audio and the emotions expressed, encompassing quite relevant information not directly related to language. The typical range for fundamental frequencies in adult males is between 80 and 160 Hz, while in adult females, it falls between 120 and 300 Hz.

2.1.4 Linear predictive coding

Officially established in 1955, linear predictive coding (or LPC) is the process by which a continuous digital signal "x", sampled at "i" discrete points in time, can be obtained as a linear combination of preceding samples [23].

$$x(i) \approx r(i) \quad = \quad \sum_{j=1}^{n} y_j x(i-j) \tag{2.3}$$

In Eq. (2.3), $r(i)$ is the estimate of the signal, n is the number of previous samples that were used to calculate it (known as the order of the signal) and y_j are the predictor coefficients, determined by minimizing the error between the actual signal and its estimate over every i sample, as shown in Eq. (2.4). In practice, it is necessary to split the signal into frames and calculate the y_j coefficients for each frame [23].

$$\sum_i [x(i) - r(i)]^2 \tag{2.4}$$

2.2 Normal Speech

NS production begins with the respiratory system, which serves as the energy source for the entire process. Akin to when we exhale, the diaphragm contracts in a controlled manner, maintaining a constant decrease in lung volume. This motion creates a controlled airflow that passes through the vocal folds in the larynx, leading to their vibration. This process is known as phonation, and it imparts a periodic component to the sound. The resulting vibrations then radiate outward from the mouth, and the auditory system perceives them as speech, identifiable by the presence of a harmonic structure.

Vowels produced in this manner are considered voiced, while consonants can be either fully or partially voiced.

2.3 Whispered Speech

WS follows a distinct production process from its normal counterpart. This difference arises because whispering lacks the vibrations of the vocal folds (and therefore the periodic component), causing the airflow to become turbulent and noisy. As a result, these turbulent characteristics make it undetectable by harmonic structures. Moreover, WS requires a greater airflow compared to NS [17]. Although it can still generate the necessary frequencies, whispered speech often leads to weakened vocal projection, reduced intelligibility, and a loss of individual sound signature.

The power spectral density is generally flatter overall, with less pronounced formants at higher frequencies [16].

2.4 Artificial Intelligence

In most recent studies of voicing decision and the eventual conversion from whispered to normal speech, AI tends to be the core component. AI can be described as the mechanism by which a computer can learn to mimic the human brain and approximate the way humans think thorough following a set of rules or algorithms.

2.4.1 Machine learning algorithms

The first type of algorithms relevant to the dissertation are dubbed machine learning algorithms (ML). A subset of AI, ML focuses on the development of programs to access data for the system's own use with the goal of providing a means to learn and improve through its own experiences without any kind of human intervention. The types of models in this algorithm utilized for the study were "logistic regression" (which serves as the baseline for the study's initial phases) (2.4.1.1), "random forest" (2.4.1.2) and "multilayer perceptron" (2.4.1.3).

2.4.1.1 Logistic regression

Logistic regression (LR) is a statistical algorithm used for binary classification. It treats inputs as independent variables and returns a probability value between 0 and 1, it can also be utilized to predict whether an instance belongs to a given class or not. LR does this by utilizing a sigmoid function to transform the continuous value output of a linear regression function into a categorical value output [5].

$$p(X;b,\omega) = \frac{1}{1+e^{-\omega.X+b}} \tag{2.5}$$

Eq. (2.5) is the final equation for LR with X being the independent input feature, b the bias term (or intercept) and ω the coefficient (or weight). Logistic Regression can be classified in three types:

- **Binomial:** The output can only be between two different dependent variables (0 or 1, Fail or Pass);
- **Multinomial:** The output can be between 3 or more unordered types of variables (different types of animals, countries);
- **Ordinal:** The output can be between 3 or more ordered types of variables (Low, Medium and High).

2.4.1.2 Random forest

To discuss the random forest (RF) model, it is necessary to first establish what ensemble learning (EL) is, and what methods of it exist.

EL is a technique that utilizes predictions of multiple models to create more accurate predictions that improve the overall performance of the learning system. Different methods of EL include:

- **Bagging (bootstrap aggregation):** The different models are trained on random subsets of data and their predictions are combined usually through averaging;

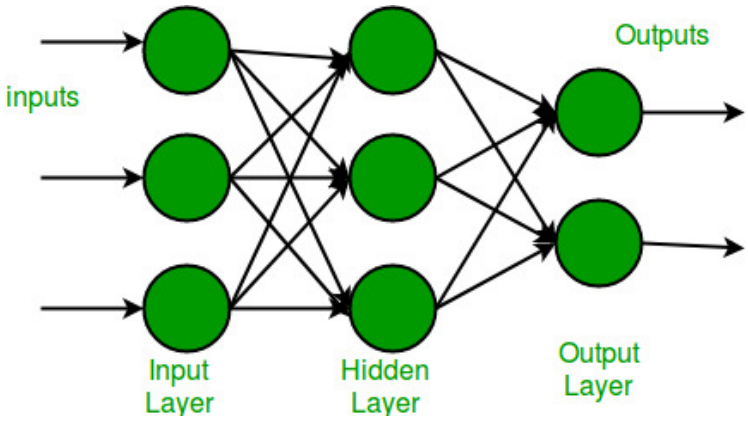


Figure 2.2: Example of a multilayer perceptron

- **Boosting:** The models are trained in a sequence, with each model focusing on the errors made by their predecessor, and the resulting predictions are combined using a weighted voting scheme;
- **Stacking:** The prediction made by a given module is used as the input for the next and this continues until the final prediction is obtained.

RF is an example of a bagging EL method, it combines the predictions of multiple decision trees and returns a more accurate one [8]. How the final result is achieved depends on the type of problem being analysed. Should it be classification, then the majority will be selected, in case of a regression problem the mean of all predictions becomes the final output.

2.4.1.3 Multilayer perceptron

A multilayer perceptron (MLP) can be described as a neural network consisting of multiple layers of nodes (also referred to as neurons) [7]. Each neuron utilizes a sigmoid activation function (2.6) that takes real value inputs and converts them into numbers between 0 and 1.

$$x = \frac{1}{(1+e^{-x})} \tag{2.6}$$

Schematic (2.2) depicts an example of a MLP. It shows the input layer as three neurons, one for each input value. Those inputs are then forwarded to the hidden layer, where they are processed. The example has only one hidden layer represented, however there can be various hidden layers present. Finally the predictions made by the hidden layer are then forwarded to a neuron each, said nodes composing the output layer [7].

2.4.2 Deep learning algorithms

The second type of algorithms are known as deep learning algorithms (DL). DL makes use of Neural Networks to mimic human brain-like behavior, especially in regards to identifying patterns

in information processing and classifying that information accordingly. They are themselves a subset of ML that handles larger sets of data and whose prediction algorithm is self-administered. The models from this algorithm chosen for the study were "convolutional neural network" (2.4.2.1) as well as two subsets of "recurrent neural network" (2.4.2.2), "long short-term memory" and "gated recurrent unit".

2.4.2.1 Convolutional neural network

Convolutional neural networks are more commonly used for computer vision tasks such as image recognition [22]. Their architecture is reminiscent of the human brain's visual processing and as such they have multiple specialized layers, each managing particular tasks:

- **Convolutional layers:** Key component responsible for feature detection and extraction;
- **Pooling layers:** Charged with downsizing the spatial dimensions of the data;
- **Fully connected layers:** Capable of making predictions based on the high-level features extracted by the preceding layers.

2.4.2.2 Recurrent neural network

A recurrent neural network (RNN) is tailored to handling sequential data such as text or time-framed information. While normally inputs received and outputs generated by a neural network are independent of the following step's, a RNN retains that information by feeding the output of a given step to its successor. The variants of RNNs studied during the dissertation were the following:

- **Long short-term memory:** Long short-term memory (LSTM) models are a direct improvement on the RNN's memory capabilities, introducing a memory cell and a gates system that controls what information the cell receives, forgets and outputs, making it excellent for tasks such as speech recognition [9];
- **Gated recurrent unit:** Much like LSTM, gated recurrent unit (GRU) models also uses a gates system to selectively update the information, albeit with a simpler structure [4].

2.5 Chapter Summary

This chapter focused on the scientific know-how required to understand the project. It tackled basic concepts for speech processing (2.1) (such as the FT, MFCCs, f_0 and LPC), the different types of speech that will be analysed (normal (2.2) and whispered (2.3)) and the various types of ML (2.4.1) and DL (2.4.2) algorithms used to achieve the results (LR, DF, MLP, CNN, LSTM and GRU).

The following chapter will go over certain selected works, with relevance to the dissertation.

Chapter 3

Related Works

This chapter will go over certain selected works, with relevance to this dissertation.

3.1 Paper Selection

The research was carried out mainly using Google Scholar, making use of its advanced search features such as filtering for specific key words or names of authors who have previously done work in the area. Some of the key words used were for example, *"candidate to voice"*, *"voiced unvoiced"* and *"context analysis"*, often accompanied by either *"whispered speech"* or *"context"* to ensure relevance. A total of 6 papers were selected using this method.

3.2 Review of Selected Papers

The papers can be sorted by the approach utilized, either rule-based, ML-based or a hybrid of both approaches.

3.2.1 Rule-based approach using spectral centroid thresholding

The study *"Glottal flow synthesis for whisper-to-speech conversion"* [20] proposes a rule-based VD method utilizing the spectral centroid of the power spectrum. It identifies broadband sounds as consonants and low-energy regions as vowels, with vocalic sounds in whispers having a spectral centroid below 4 kHz and unvoiced consonants above this threshold. To differentiate between vowels and consonants, a voiced/unvoiced frequency threshold is established, where a centroid below this threshold sets the VD value to 0, and above it to 1. To address slow VD transitions caused by gradual centroid changes, a non-linear sigmoid function mapping is proposed. The method's objective evaluation resulted in error rates of 7.6% for voiced sounds, 10.1% for unvoiced sounds, and a total error rate of 17.7%.

Critical Analysis: This paper presents a novel approach to enhancing the naturalness of speech conversion by focusing on the synthesis of a phonated source from whisper parameters. The proposed method shows promising improvements over traditional techniques, especially in

terms of naturalness and intelligibility. However, providing more detailed information on the dataset, computational performance, phoneme category handling, and ensuring reproducibility would strengthen the overall impact and applicability of the research. Expanding the comparative analysis to include more baseline methods and state-of-the-art approaches would further substantiate the method's efficacy.

3.2.2 Rule-based approach combining context analysis to different types of thresholding

The document *"Improvement of voicing decisions by use of context"* [15] discusses enhancing voicing decisions in speech compression and recognition by incorporating contextual information from previous segments. Traditional methods consider each segment independently, but this paper proposes using the voicing status of the previous segment to improve accuracy. By applying two thresholds based on the previous segment's voicing status, error rates are significantly reduced. For instance, using low-frequency energy, a typical improvement is a 15% drop in error rate. Experiments demonstrate the effectiveness of this approach: a double threshold reduced errors in low-frequency energy from 36 to 28 in 2410 segments, improved voicing accuracy by 17% using the energy ratio, and reduced zero-crossing count errors by 14%. The paper concludes that incorporating context is a simple yet effective method to improve speech processing performance that while specific thresholds may not be optimal, the contextual approach is robust and can be further enhanced with more sophisticated techniques, such as dynamic programming, though these may require more computational resources.

Critical Analysis: The study introduces the concept of improving VDs by incorporating minimal contextual information. The method shows promising results in classification performance with minimal computational overhead. However, the analysis is limited by a lack of detailed data description, broader comparative analysis, and comprehensive performance metrics. Enhancing these aspects and including ML techniques for comparison could further validate and potentially improve the proposed approach's effectiveness and relevance.

3.2.3 LSTM-based approach applied to DNN

The paper *"LSTM-based robust voicing decision applied to DNN-based speech synthesis"* [21] (with DNN meaning deep neural network) presents a novel approach for improving VDs in statistical parametric speech synthesis using a LSTM network. Accurate voiced/unvoiced classification is crucial for high-quality speech generation, and traditional methods like RAPT, REAPER, and STRAIGHT struggle with creaky or low amplitude speech. The proposed method uses the Fast Fourier Transform (FFT) to compute the power spectrum for each speech frame and trains an LSTM network on these features for classification. This significantly reduces misclassification rates, enhancing SPSS speech quality. The LSTM network, trained on the Keele database, demonstrated superior performance on the CMU Arctic, Keele, and MIR-1K databases compared to

traditional techniques. Integrating this method into a DNN-based SPSS framework (SPSS meaning Statistical Parametric Speech Synthesis) using the HTS toolkit improved speech quality, as confirmed by subjective evaluations. The study concludes that improved voiced/unvoiced classification enhances speech synthesis quality, with future work potentially exploring dimensionality reduction for further optimization.

Critical Analysis: The method proposed by the study demonstrates superior performance compared to traditional techniques and has the potential to enhance speech synthesis quality. However, to strengthen the work, the authors should provide more detailed information on the data preprocessing, training procedures, computational performance, and reproducibility aspects. Addressing these areas would make the research more accessible and verifiable for the broader scientific community.

3.2.4 DNN-based approach, using MFCCs as training parameters

The document *"Real-time whispered to natural mode conversion of audio signals"* [14] discusses an approach to converting whispered speech into naturally phonated speech, aiding laryngectomy patients who can only produce whispered sounds. The authors emphasize the importance of this conversion for effective communication, particularly in privacy-sensitive situations like telephony. The method employs two feed-forward DNNs to classify voiced and unvoiced segments and predict pitch, enhancing the naturalness and robustness of the conversion. This approach surpasses existing algorithms that lack training or produce robotic-sounding outputs. The study highlights the prosodic information retained in WS, reviews the limitations of current medical solutions, and categorizes existing whisper-to-normal conversion methods. The methodology includes a "whisperize" module for aligning whispered and normal audio samples and details the DNN architecture for pitch prediction and segment classification. Results show that the method generates natural-sounding speech from whispered input, with potential applications in medical fields and everyday communication, suggesting the feasibility of real-time conversion systems to improve speech intelligibility.

Critical Analysis: The authors proposed two feed-forward DNNs for predicting pitch and classifying voiced/unvoiced segments, claiming improved robustness and naturalness in the converted speech. However, the paper lacks any mention of quantitative performance metrics, computational benchmarks, and detailed reproducibility information, necessitating further research to validate and compare their method comprehensively.

3.2.5 GMM-based approach to NAM-to-speech conversion

The paper *"Predicting f_0 and voicing from NAM-captured whispered speech"* [24] (NAM meaning non audible murmur) proposes a method to improve the conversion of non-audible murmurs into audible speech, focusing on WS. Critiquing the limitations of gaussian mixture model (GMM) methods for estimating f_0 and voicing, the authors propose a new approach that separates these processes. They use a feed-forward neural network to identify voiced segments in WS and a GMM

for estimating the melodic contour of these segments. By aligning whispered and normal speech utterances during training the conversion accuracy is enhanced. Experimental evaluations show that the proposed method significantly reduces error rates in VDs and improves f_0 estimation, resulting in better intelligibility and naturalness of the converted speech, as confirmed by listener tests. The study concludes that this approach not only improves speech quality from whispered inputs but also enhances NAM-to-speech conversion technologies.

Critical Analysis: The described method gives insight on a training process involving 200 WS utterance pairs from a French male speaker, with the new method achieving a 6.8% error rate in voicing detection compared to 9.2% for the GMM-based system, and better f_0 estimation, intelligibility, and naturalness in converted speech. Despite its promising results, the study lacks comparisons to other advanced techniques, necessitating further research for comprehensive evaluation and reproducibility.

3.2.6 Hybrid approach using a KNN phoneme classifier followed by rule-based voicing decision using spectral centroid thresholding

The study *"Voicing decision based on phonemes classification and spectral moments for whisper-to-speech conversion"* [2] describes a low-resource VD system suitable for real-time applications. This system classifies WS frames into phoneme classes using spectral centroid and spread via the k-nearest neighbors (KNN) algorithm and then discriminates voiced phonemes from unvoiced ones based on class-dependent spectral centroid thresholds. Trained with an in-house database, the proposed approach is compared to a simpler single centroid threshold method, achieving VD accuracy above 91% and reducing systematic VD errors, thereby facilitating users' real-time speech adaptation to compensate for residual VD errors.

Critical Analysis: Overall, the paper presents a unique approach to the VD in speech conversion, leveraging spectral moments for phoneme classification. It achieves high classification accuracy and shows potential for real-time application. However, additional details on data characteristics, computational performance, broader comparisons, and reproducibility would strengthen the study's impact and utility in the field.

3.2.7 DL-based approaches to WS segmentation

The dissertation report *"Whispered speech segmentation based on deep learning"* [16] outlines a VD involving the development of a subsystem that accurately distinguishes between phonemes that are candidates for voicing and those that are not, which is essential for whispered-to-normal speech conversion systems. The study employs advanced DL models, specifically temporal convolutional networks, CNN, LSTM, GRU, and Transformers, to enhance the accuracy of voicing decisions. A comparative analysis is conducted using two feature subsets: a baseline set of 49 MFCCs and a refined subset selected through feature engineering. The results demonstrate that the temporal convolutional network (TCN) model, when applied with the selected feature subset, significantly outperforms other models in various performance metrics, achieving high accuracy

and precision, which underscores the potential of deep learning in improving the intelligibility and quality of speech for individuals relying on whispered communication.

Critical Analysis: The author provides a detailed description of the dataset as well as accurately describes the VD process and DL model structure utilized, leading to high reproducibility. While he does mention the fact that the DL models chosen have higher computational requirements and a system with matching capabilities might be able to achieve better results, they are still quite decent reaching values of over 90% in the classification metrics.

3.3 Chapter Summary

This chapter went over selected works (3.1), with relevance to the project, describing their approaches to VD and analysing them accordingly (3.2).

The next section will go into detail on the steps and procedures undertaken during the development of the study.

Chapter 4

Methodology

This chapter goes into detail on the steps and procedures undertaken during the development of the study. It also gives information on the conditions of the tests, specifics on the dataset used to create the predictions, and the metrics by which said predictions were evaluated.

4.1 Hardware and Software

Here a detailed list of hardware (4.1.1) and software (4.1.2) specifications is presented for the device used to conduct the tests:

4.1.1 Hardware

- **Central Processing Unit (CPU):** 12th Gen Intel(R) Core(TM) i5-12450H;
- **Graphics Processing Unit (GPU):** NVIDIA GeForce RTX 3050;
- **Solid State Drive (SSD):** NVMe Micron MTFDHBA512QFD;
- **Random Access Memory (RAM):** 16 GB.

4.1.2 Software

- **Operating System (OS):** Windows 11 Home Version 23H2 Build 22631.3880;
- **Programming Languages:** Python 3.12 [26];
- **Python Packages:** Librosa 0.10.2.post1 [12]; Numpy 1.26.4 [6]; Pandas 2.2.1 [13]; Pickle 4.0 [25]; Plotly 5.19 [10]; Scikit-learn 1.4.1.post1 [19]; Tensorflow 2.16.1 [1]; Librosa
- **GPU Driver:** NVIDIA driver 32.0.15.6070.

4.2 Specifics on the Dataset

The dataset used for this study was constructed via the recordings of a few select individuals who were properly trained to ensure a clear and leveled intonation of the necessary NS and WS segments.

4.2.1 Selection process and recording environment

A total of 17 individuals were selected, 9 males and 8 females, with ages between 22 and 33 years old from Aveiro and Coimbra. These individuals were required to have no vocal or respiratory impairments on the day of recording as well no previous history of being afflicted with them or having had surgery done on their larynx, they had to be able to produce all the needed vocal tasks without need of orthodontic devices and have European Portuguese as their first language.

The recordings were conducted in a quiet setting, with background noise only reaching around 15.1 *dB LAeq* (LAeq being the A-weighted, equivalent continuous sound level in decibels, having the same total sound energy as the fluctuating level measured). The participants utilized the *Sennheiser Ear Set 1* microphone worn on their heads and the recordings were sampled at a rate of 48,000 *Hz* with a 16 *bit* resolution.

4.2.2 Dataset structure

The working dataset is comprised of 27 files for NS and WS each, a total of 54 per participant. The materials that composed the corpus included 4 sustained sibilants, 4 sustained oral vowels, 12 disyllabic words, 6 Consensus Auditory-Perceptual Evaluation of Voice sentences [11], and an extract of the phonetically balanced text, *"The North Wind and the Sun"* containing 98 words and 196 syllables [18].

4.3 Dataset Preprocessing

Before being able to be analysed and used to make predictions, the dataset needed to be refined by undergoing certain procedures made to improve the quality of the information and ease its processing. This stage of the study is called the preprocessing.

4.3.1 Downsampling

The first step was to downsample the audio files. While originally they were sampled at 48,000 *Hz* (4.2.1), by resampling them to 22,050 *Hz* the results showed increased efficiency in the analysis while still preserving the most relevant spectral content.

4.3.2 Phone-based annotation and segmentation

After the downsampling of the audio, the next step was to divide the dataset into phone segments and arrange them in a table to simplify the analysis:

- **Sex:** The speaker's gender;
- **Speaker Identification:** The speaker's unique ID;
- **Task:** Which task is associated to the segment;
- **Speech Mode:** Whether it is NS or WS;
- **Sequence Index:** The phone's position within the task;
- **Segment's Waveform Audio File Format File:** A WAV file containing samples of the audio track;
- **Sampling Frequency:** The sampling frequency of the audio track;
- **Phonetic Annotation Label:** The PAL for the segment.

4.3.3 Selection and cleaning

To ensure more reliable PALs the dataset needs to undergo a selection and cleaning steps. The selection was carried out according to the two criteria in Eq. (4.1) and Eq. (4.2).

$$(PAL = silence) \vee (length(PAL) = 3 \wedge PAL[-1] \in \text{"}w\text{"}, \text{"}W\text{"}) \tag{4.1}$$

$$SM = \text{"02"} \tag{4.2}$$

The first condition (4.1) states that the entry PAL must either be a "*silence*" or a character with length 3 ending in either "*w*" or "*W*", the second condition (4.2), filters the entries by speech mode ensuring only those of WS. Combined, these conditions eliminate entries with unreliable PALs from the table.

The cleaning is a simpler process, consisting of replacing "-" and "w" characters with "_" and "W" characters respectively.

4.3.4 Labelling segments as voicing candidates

By adding a new boolean value called candidate to voicing (CTV), the entries of the table can now be classified on whether they are CTV or not candidate to voicing (NCTV), depending on their Speech Assessment Methods Phonetic Alphabet (SAMPA) labels. The results are illustrated on table 4.1, where those considered to be CTV have that value set to 1, while those who are considered to be NCTV have it set to 0.

Table 4.1: SAMPA phonetic annotation labels of segments and candidate to voicing label.

Phonetic Annotation table	**SAMPA Symbol**	**CTV/NCTV**
1_W	1	1
4_W	4	1
6_W	6	1
A_W	A	1
E_W	E	1
L_W	L	1
N_W	N	1
O_W	O	1
R_W	R	1
Z_W	Z	1
a_W	a	1
b_W	b	1
d_W	d	1
e_W	e	1
g_W	g	1
i_W	i	1
l_W	l	1
m_W	m	1
n_W	n	1
o_W	o	1
u_W	u	1
v_W	v	1
z_W	z	1
S_W	S	0
f_W	f	0
k_W	k	0
p_W	p	0
s_W	s	0
silence	sil	0

4.3.5 Normalization of audio segments

The SWAV column of the dataset was subjected to normalization, employing a standardization technique.

$$SWAV_{Normalized} = \frac{SWAV_{Original} - \mu}{\sigma} \tag{4.3}$$

In Eq. (4.3), $\text{SWAV}_{Normalized}$ is the standardized SWAV, $\text{SWAV}_{Original}$ is the pre-standardized version, μ is the mean and σ is the standard deviation of all $\text{SWAV}_{Original}$. Through its application we can remove the bias that originates from the differing scales of the original SWAV.

4.4 Feature Engineering

Feature engineering is an important phase in the analysis of the dataset as it is what transforms the speech data into a set of meaningful and representative features to be used for the modeling phase that proceeds it.

4.4.1 Feature extraction

To perform the feature extraction on the preprocessed dataset the Librosa [12] package was used. It made use of each entry's phone's normalized WAV files and sampling frequencies to compute several features. This operation adopted a Frame Size (FS) of 1,024 samples and a Hop Size (HS) of 512 samples, which are considered the default values for all remaining parameters unless stated otherwise.

While several features can be extracted via this method, the only one used for the purposes of this study were the 49 MFCCs, that provide a 49-dimensional vector for each frame.

4.4.2 Feature normalization

Following their extraction, the MFCCs were subjected to the same normalization process as the SWAVs (4.3.5) with the goal of removing bias in future analysis.

$$MFCC_{Normalized} = \frac{MFCC_{Original} - \mu}{\sigma} \tag{4.4}$$

In Eq. (4.4), $\text{MFCC}_{Normalized}$ is the resulting standardized MFCC values, $\text{MFCC}_{Original}$ is the pre-standardized version, μ is the mean and σ is the standard deviation of all $\text{MFCC}_{Original}$.

4.4.3 Dataset explosion from segments to frames

The current dataset consisted of several entries, each corresponding to a phone and respective array of MFCCs. These values were determined for each frame using a FS of 1,024 samples, with centers located at every HS of 512 samples. However, the most ideal structure for future analysis would be one where each entry corresponded to the MFCCs extracted from a single frame. By making use of the explosion technique (where a large scale of data is rapidly generated and stored), and applying it to the dataframe based on the respective MFCCs, each resulting outputted entry retained the corresponding attribute values from its original state (ensuring their connection), while also making the data more easily interpretable by ML and DL algorithms.

4.4.4 Class rebalancing through selective silence frame reduction

An analysis on the dataset's CTV and NCTV classes distribution was conducted post explosion process, revealing an imbalance in the NCTV class that could potentially interfere with the learning process and lower its effectiveness. This imbalance was caused due to the class having an

abundance of silence frames, so as a means of correcting it, a 2-step method for selectively reducing the number of silence frames was established:

1. Identify periods of continuous silence frames;
2. Iteractively remove a percentage of the silence frames from the middle of those segments.

This method serves as both a way to facilitate class distribution in the dataset, as well as preserving the frames near the segment boundaries for they might prove critical for a context based analysis.

4.4.5 Context-based analysis

Context can be described as the previous frames that have already gone through the classification process and therefore are already identified as either voiced or unvoiced. The model infrastructure for binary classification used in this study, utilized context to improve their results by means of a sliding window with adjustable sizes, defining the amount of context frames that the analysis should employ.

4.4.6 Baseline definition

For the purpose of the study one of the ML or DL models was chosen to act as the baseline.

4.5 Evaluation Metrics Definition

To evaluate the models chosen for the study, a total of 5 performance metrics were chosen, specifically Accuracy (4.5.1), Precision (4.5.2), F1 Score (4.5.4), Recall (4.5.3) and AUC (4.5.5).

4.5.1 Accuracy

Accuracy quantifies the proportion of true predictions out of the total number of predictions generated by the model, and can be exemplified by Eq. (4.5).

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives} \tag{4.5}$$

4.5.2 Precision

Precision, as defined in Eq. (4.6), quantifies the proportion of true positive predictions out of the total number of positive predictions generated by the model, demonstrating how precise it is at making positive predictions.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4.6}$$

4.5.3 Recall

Recall or sensitivity is the rate of positive instances that were correctly identified, as defined in Eq. (4.7).

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{4.7}$$

4.5.4 F1 Score

F1 Score, is the harmonic mean of Precision and Recall, providing a balance between the two metrics and can be calculated using Eq. (4.8).

$$F1Score = 2.\frac{Precision.Recall}{Precision + Recall} \tag{4.8}$$

4.5.5 Area Under the ROC Curve

AUC returns a measure of the model performance across all possible classifications, despite not having a formula like the rest it is calculated by plotting the recall over the false positive rate (this one being shown in Eq. (4.9)).

$$FalsePositiveRate = 1 - \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{4.9}$$

4.6 Assessment of the Best Model

To determine which model performed best in the dataset, an assessment of all the models needed to be carried out. The method chosen for evaluation was a train-test split (TTS) carried out a total of 10 times per context window size. The TTS segments were divided in a 70% ratio for the training set, 15% ratio for the validation set, and 15% ratio to the testing set. Once the model was finished with the testing phase it would output the performance metrics. The Accuracy and Precision metrics would determine which model was best suited for context-base VD.

4.7 Chapter Summary

This chapter went into detail on the steps and procedures undertaken during the development of the study. It gave information on the conditions of the tests, most specifically the hardware and software used to run it (4.1), details on the dataset used to create the predictions, mainly how it was created (4.2) and preprocessed (4.3), and the metrics by which said predictions were evaluated (4.4).

The next chapter will describe the results achieved in the study.

Chapter 5

Results and Discussion

This chapter covers the results obtained by the study, going over the dataset aquisition and its preprocessing, feature selection, assessment and comparison of all models' performance to determine which model is the best and what context interval returns the best results.

5.1 Dataset Aquisition

Applying the method detailed in "Methodology" Chapter, in the "Specifics on the Dataset" Section (4.2), resulted in the creation of a dataset comprised of 54 audio files (half NS, half WS) per each of the 17 participants.

5.2 Dataset Preprocessing (Results)

Making use of the methods in Section "Data Preprocessing" of the "Methodology" Chapter (4.3), the previously phonetically annotated dataset was transformed into a tabular dataset with 12,718 entries that each possessed the following characteristics:

- **Sex:** The speaker's gender;
- **Speaker Identification:** The speaker's unique ID;
- **Task:** Which task is associated to the segment;
- **Speech Mode:** Whether it is NS or WS;
- **Sequence Index:** The phone's position within the task;
- **Segment's Waveform Audio File Format File:** A WAV file containing samples of the audio track;
- **Sampling Frequency:** The sampling frequency of the audio track, which is 22,050 Hz post downsampling;

- **Phonetic Annotation Label:** The PAL for the segment;
- **CTV Label:** Boolean label that indicates if the segment is a CTV or not.

5.3 Feature Engineering (Results)

This section details the results from the processes of feature engineering described in the "Feature Engineering" Section of the "Methodology" Chapter (4.4), more specifically the feature extraction (5.3.1), feature normalization (5.3.2), dataset explosion from segments to frames (5.3.3), class rebalancing through selective silence frames reduction (5.3.4), context-based analysis (5.3.5) and baseline definition (5.3.6).

5.3.1 Feature extraction

The method from Section "Feature Extraction" in the "Methodology" Chapter (4.4.1) resulted in the same tabular dataset from the previous Section (5.3), with the addition of the MFCCs.

5.3.2 Feature normalization

Using the method described in Subsection "Feature Normalization" from the "Methodology" Chapter (4.4.2), the MFCCs in the tabular dataset were normalized.

5.3.3 Dataset explosion from segments to frames

The method described in Subsection "Dataset explosion from segments to frames" of Chapter "Methodology" (4.4.3) resulted in an exploded version of the tabular WS dataset with normalized MFCCs which had the following characteristics:

- 241,777 entries;
- Each entry corresponds to an audio frame with FS 1024 and HS 512;
- Each entry retains the corresponding attribute values from the original entry, preserving that connection between the two.

5.3.4 Class rebalancing through selective silence frame reduction

After analysing the dataset, a class distribution was detected due to an overrepresentation of NCTV entries (163,262 out of 241,777). To ammend this, the method described in Subsection "Class rebalancing through selective silence frame reduction" of Chapter "Methodology" (4.4.4) was used, resulting in the amount of entries for that class to be reduced to around 48% of its original value.

The resulting values for the classes were:

- 78,515 CTV entries;

- 78,513 NCTV entries;
- 157,028 entries total.

5.3.5 Context-based analysis

To allow a more beneficial context-based analysis as described in Chapter "Methodology", in particular the "Context-based analysis" Subsection (4.4.5), it would be more beneficial to first prove that context does indeed have an effect on the model's performance. For that reason, the first batch of tests were conducted without context being a factor, the second batch of predictions would show the effect context had, and the third sequence of experiments would attempt to discover if there was a particular context size or interval of sizes where the models performed at their peak.

5.3.6 Baseline definition

Considering the nature of the evaluation process, the Logistic Regression model 2.4.1.1 was chosen as the baseline.

5.4 Assessment of the Best Model (Results)

Listed in this section are the results of the procedures described in the "Assessment of the best model" Section of the "Methodology" Chapter (4.6), in accordance with the intended context analysis (5.3.5).

5.4.1 Results without context

10 iterations of TTS evaluation were conducted in the established ML and DL models: LR (acts as the baseline), DF, MLP, CNN, LSTM and GRU. The resulting performance metrics and their respective standard deviations are listed in Table 5.1 and Figure 5.1, including:

- **Accuracy:**
  - The RF model tops the list with an Accuracy of $85.79\% \pm 0.14\%$;
  - It is followed by the LSTM model ($85.32\% \pm 0.18\%$), GRU model ($85.05\% \pm 0.13\%$) and CNN model ($85.04\% \pm 0.02\%$);
  - The MLP model has an Accuracy of $84.22\% \pm 0.14\%$;
  - The LR model ranks last with an Accuracy of $81.52\% \pm 0.13\%$.
- **Precision:**
  - The CNN model tops the list with a Precision of $85.74\% \pm 0.87\%$;
  - It is closely followed by the LSTM model with a Precision of $85.54\% \pm 0.57\%$;

- Then followed by the GRU model (84.36% ± 0.20%), the MLP model (84.26% ± 0.15%) and the RF model (83.85% ± 0.30%);
- The LR model ranks last with a Precision of 81.34% ± 0.26%.

- **Recall:**
  - The RF model tops the list again, by a large margin, with a Recall of 88.71% ± 0.05%;
  - It is followed by the GRU model (86.11% ± 0.20%), and then the LSTM model (85.03% ± 0.72%);
  - Then followed by the MLP model (84.22% ± 0.31%) and the CNN model (84.13% ± 1.33%);
  - The LR model ranks last with a Recall of 81.79% ± 0.17%.

- **F1 Score:**
  - Again, the RF model tops the list with an F1 Score of 86.21% ± 0.18%;
  - It is followed by the LSTM model (85.28% ± 0.22%), GRU model (85.22% ± 0.20%) and CNN model (84.92% ± 0.25%);
  - The MLP model has an F1 Score of 84.24% ± 0.23%;
  - The LR model ranks last with an F1 Score of 81.56% ± 0.14%.

- **AUC:**
  - Once more the RF model tops the list with an AUC of 93.77% ± 0.13%;
  - It is followed by the LSTM model (93.49% ± 0.11%), CNN model (93.42% ± 0.001%) and GRU model (93.32% ± 0.14%);
  - The MLP model has an AUC of 92.45% ± 0.21%;
  - The LR model ranks last with an AUC of 89.92% ± 0.11%.

By analysing these results, it becomes clear that the RF model is the most promising in all respects except for precision. Results for the LSTM, GRU and CNN models tend to be within small margin of each other and all models seem to have far better results when compared to the baseline.

Table 5.1: Classifiers' performance metrics and standard deviations obtained from 10 iterations of TTS evaluation without context.

Model	Accuracy	Precision	Recall	F1 Score	AUC
LR	81.52% ± 0.13%	81.34% ± 0.26%	81.79% ± 0.17%	81.56% ± 0.14%	89.92% ± 0.11%
RF	85.79% ± 0.14%	83.85% ± 0.30%	88.71% ± 0.05%	86.21% ± 0.18%	93.77% ± 0.13%
MLP	84.22% ± 0.14%	84.26% ± 0.15%	84.22% ± 0.31%	84.24% ± 0.23%	92.45% ± 0.21%
CNN	85.04% ± 0.02%	85.74% ± 0.87%	84.13% ± 1.33%	84.92% ± 0.25%	93.42% ± 0.001%
LSTM	85.32% ± 0.18%	85.54% ± 0.57%	85.03% ± 0.72%	85.28% ± 0.22%	93.49% ± 0.11%
GRU	85.05% ± 0.13%	84.36% ± 0.20%	86.11% ± 0.20%	85.22% ± 0.20%	93.32% ± 0.14%

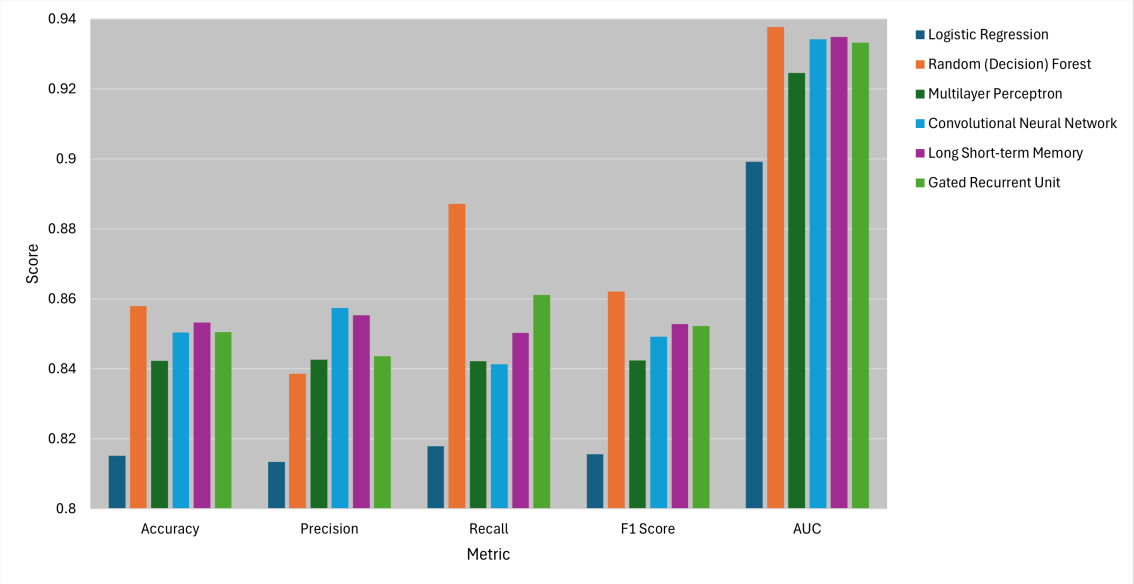


Figure 5.1: Classifiers' performance metrics obtained from 10 iterations of TTS evaluation without context.

5.4.2 Results with context

Another 10 iterations of TTS evaluation were conducted in the established ML and DL models, this time with context. The context size utilized for this study was 29 *ms*, estimated to be the approximate size of a word in frames [16]. The resulting performance metrics and their respective standard deviations are listed in Table 5.2 and Figure 5.2, including:

- **Accuracy:**
  - The LSTM model tops the list with an Accuracy of 95.93% ± 0.17%;

- It is closely followed by the GRU model with an Accuracy of $95.37\% \pm 0.08\%$;
- Then followed by the CNN model ($89.36\% \pm 0.33\%$), the MLP model ($87.93\% \pm 0.23\%$) and the RF model ($87.24\% \pm 0.19\%$);
- The LR model ranks last with an Accuracy of $82.76\% \pm 0.19\%$.

- **Precision:**
  - Again, the LSTM model tops the list with a Precision of $95.92\% \pm 0.37\%$;
  - It is closely followed yet again by the GRU model with a Precision of $95.12\% \pm 0.32\%$;
  - Next we have the CNN model ($90.64\% \pm 1.07\%$), the MLP model ($87.48\% \pm 1.04\%$) and the RF model ($85.91\% \pm 0.44\%$);
  - The LR model performs the lowest with a Precision of $82.50\% \pm 0.18\%$.
- **Recall:**
  - The LSTM model continues at the top of the list with a Recall of $95.94\% \pm 0.41\%$;
  - It is again followed by the GRU model with a Recall of $95.67\% \pm 0.35\%$;
  - Then followed by the RF model ($89.08\% \pm 0.06\%$), MLP model ($88.53\% \pm 0.64\%$) and CNN model ($87.79\% \pm 0.35\%$);
  - The LR model ranks lowest with a F1 Score of $83.13\% \pm 0.38\%$.
- **F1 Score:**
  - Once more, the LSTM model tops the list with an F1 Score of $95.93\% \pm 0.17\%$;
  - It is once again followed by the GRU model with its F1 Score of $95.39\% \pm 0.01\%$;
  - Then followed by the CNN model, MLP model and RF model with F1 Score of ($89.19\% \pm 0.34\%$), ($88\% \pm 0.21\%$) and ($87.47\% \pm 0.26\%$) respectively;
  - The LR model once again performs the lowest with an F1 Score of $82.81\% \pm 0.2\%$.
- **AUC:**
  - The LSTM model has the highest value of AUC, it being $99.31\% \pm 0.06\%$;
  - The GRU model is next with AUC equal to $99.17\% \pm 0.35\%$;
  - Then by the CNN model ($96.03\% \pm 0.14\%$), RF model ($95.09\% \pm 0.03\%$) and MLP model ($95.04\% \pm 0.05\%$);
  - The LR model remains at the bottom with AUC of $91.43\% \pm 0.22\%$.

By analysing these results, the impact of context is reflected on the overall increase of the models' effectiveness, the LSTM and GRU models, in particular, show a noticeable gap compared to the other models. In accordance to the evaluation criteria established, the LSTM model, when making predictions in context, is the best suited for VD, this could be due to the fact that the LSTM

Table 5.2: Classifiers' performance metrics and standard deviations obtained from 10 iterations of TTS evaluation with context.

Model	Accuracy	Precision	Recall	F1 Score	AUC
LR	82.76% ± 0.19%	82.50% ± 0.18%	83.13% ± 0.38%	82.81% ± 0.2%	91.43% ± 0.22%
RF	87.24% ± 0.19%	85.91% ± 0.44%	89.08% ± 0.06%	87.47% ± 0.26%	95.09% ± 0.03%
MLP	87.93% ± 0.23%	87.48% ± 1.04%	88.53% ± 0.64%	88% ± 0.21%	95.04% ± 0.05%
CNN	89.36% ± 0.33%	90.64% ± 1.07%	87.79% ± 0.35%	89.19% ± 0.34%	96.03% ± 0.14%
LSTM	95.93% ± 0.17%	95.92% ± 0.37%	95.94% ± 0.41%	95.93% ± 0.17%	99.31% ± 0.06%
GRU	95.37% ± 0.08%	95.12% ± 0.32%	95.67% ± 0.35%	95.39% ± 0.01%	99.17% ± 0.05%

model as well as the GRU models are offsets of RNN, whose purpose is to make predictions using past data (akin to the context), so it would have a suitable framework for this purpose.

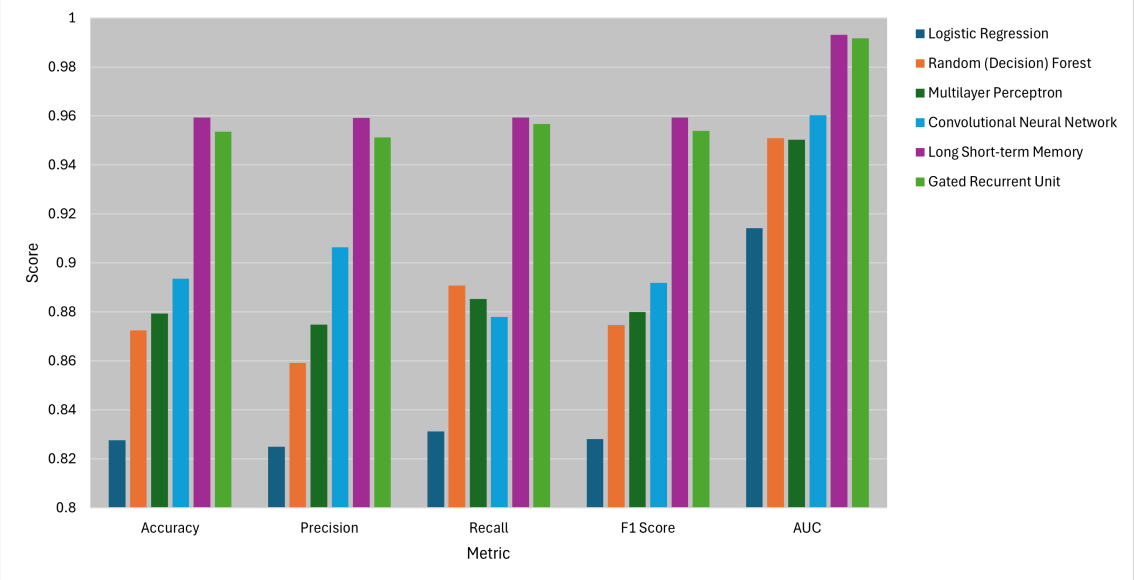


Figure 5.2: Classifiers' performance metrics obtained from 10 iterations of TTS evaluation with context.

5.4.3 Analysis of context intervals

This final batch of tests were performed using only the best model, as it was the one who demonstrated being most sensitive to the presence and effect of context. The conditions of these trials still consisted of 10 iterations of TTS evaluation, however unlike with the previous experiments, these saw the use of the sliding window feature to perform VD predictions with different values for the context.

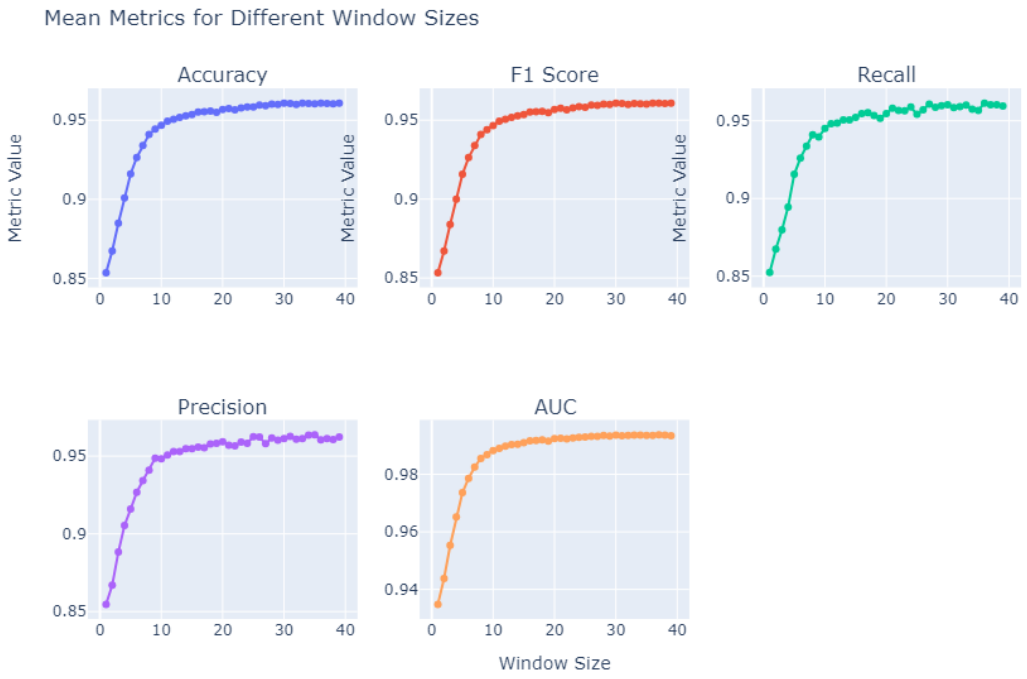


Figure 5.3: Classifiers' performance metrics obtained from 10 iterations of TTS evaluation with sliding context window.

The chosen interval for the sliding window was from 1ms to 40ms, the predictions were realized, and the metrics were calculated per size of the context window. Figure 5.3 shows the results in graphical form, where each of the points corresponds to the calculated metric value.

The graph in Figure 5.3 shows that, in an initial phase, (around 1 to 10 *ms*) the effectiveness of the models greatly increases. However after that section, the effectiveness begins to stagnate and at around 29 to 30 *ms* the increments are barely indistinguishable. This suggests some degree of hysteresis around window size 30, implying that regardless of how much context is supplied to the model from that point onward, the performance metrics will not see any improvement.

5.5 Chapter Summary

This chapter covered the results obtained by the study, going over the dataset acquisition (5.1) and its preprocessing (5.2), feature selection (5.3), assessment and comparison of all models' performance to determine which model is the best (5.4), comparison between models trained without context (5.4.1), and with it (5.4.2), as well as what context interval returns the best results (5.4.3).

The following chapter is the final one, and will cover what conclusions were reached as well as what plans are envisioned for the future.

Chapter 6

Conclusions and Future Work

The following chapter will cover what conclusions were reached as well as what plans are envisioned for the future.

6.1 Summary of Conclusions

This section provides a summary of what conclusions were reached from this study and the steps taken to complete it, namely the preprocessing of the dataset (6.1.1), feature engineering (6.1.2), assessment and comparison of the models to determine which performed best (6.1.3) and analysis of the context intervals (6.1.4).

6.1.1 Data preprocessing

The preprocessing steps the dataset was subjected to included downsampling of audio files, their segmentation and reorganization into a table. From that table's entries only a few specific ones were chosen to continue the process after being cleaned according to particular criteria. The segments were defined as either CTV or NCTV based on phonetic annotations and then standardized.

6.1.2 Feature engineering

Feature engineering focused on the process of extracting the MFCCs, which were promptly standardized. Frames with size 1,024 and with a HS of 512 samples were created for the MFCC feature. The dataset was subjected to a rebalancing with the goal of deleting a selection of silence frames due to their overrepresentation. The process for conducting context-based analysis was discussed and the requirements for the baseline.

6.1.3 Assessment of the best performing model

The ML and DL models (LR, RF, MLP, CNN, LSTM and GRU) were trained by means of a TTS approach, the performance metrics were acquired for evaluation. Two comparisons were

made, the first between models with LR as the baseline, and the second between the set of predictions obtained using context and the one obtained without context. It was concluded that the best performing model was LSTM whose predictions were made using context, with an Accuracy of $95.93\% \pm 0.17\%$, Precision of $95.92\% \pm 0.37\%$, Recall of $95.94\% \pm 0.41\%$, F1 Score of $95.93\% \pm 0.17\%$, and AUC of $99.31\% \pm 0.06\%$.

6.1.4 Analysis of context intervals

A final set of predictions was created, with the models trained by means of a TTS approach with context, however the window size for said context was an interval of 1 to 40 *ms*. It was concluded by analysing the graph in Figure 5.3 that while there is a period of fast improvement at the start (1 to 10 *ms*) the values tend to settle around 95% effectiveness with a window size of 30 *ms*.

6.2 Fulfillment of the Intended Goals

The specific goals to achieve in this study were the following:

1. **Description of a phonetically annotated speech dataset**;
2. **Preprocessing of the dataset**;
3. **Feature engineering**;
4. **Evaluation metrics definition**;
5. **Assessment of the best performing model**;
6. **Analysis of context intervals**.

In short, all the goals were achieved in their entirety albeit one with a different outcome than expected. The study gives details on a phonetically annotated speech dataset, and performs the necessary preprocessing steps to utilize it correctly. The feature engineering process was successful, the MFCCs were extracted correctly and the evaluation metrics for the models were capable of providing the necessary information to compare and assess them. The initial theory for context window sizes was that there existed an interval where the models performed best, however instead, results showed a point where the context no longer influenced the results due to a hysteresis effect.

6.3 Contributions and Limitations

This study has the potential to make ground regarding the subject of VD. It demonstrated the positive effects that context can have on the classification process of the models, which in turn can lead to more natural sounding converted speech. The assessment of LSTM as the most well suited model for context based VD approaches also contributes to an increase in usage of that particular DL model in similar tasks.

Like most works, there were several limiting factors in this study that prevented a full solution to the problem at hand. First the lack of greater computational resources and the consequential increase in time consumption were the primary factors behind why more extensive tests regarding context above a window of size 40 couldn't be realized as well as a consistent monitoring of the train time and test time. Besides that, there was the concept of the dataset being too restricting in its criteria for selection considering the multitude of different cases in the real world, or that only six ML and DL models were analysed when there exist countless others.

6.3.1 Future work

Regarding the limitations described in the previous section (6.3), future studies made on this topic should aim to address those issues. For example performing the experiments on a system with more RAM or a better GPU, a broader selection process for the dataset and an expanded list of ML and DL models for analysis.

6.4 Chapter Summary

This chapter covered what conclusions were reached, summarizing the results (6.1). It also restated what the goals for the study were and if they were accomplished or not (6.2), finally ending with the contributions offered by this dissertation as well as what limited it (6.3), and what future works of a similar nature should keep in mind to keep improving the knowledge of the topic (6.3.1).

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Luc Ardaillon, Nathalie Henrich Bernardoni, and Olivier Perrotin. Voicing decision based on phonemes classification and spectral moments for whisper-to-speech conversion. In *Interspeech 2022 - 23rd Annual Conference of the International Speech Communication Association*. ISCA, 2022.

[3] Michael Brock. Data visualization using the fourier transform, June 2019. Insight, Inc., available in https://insightincmiami.org/blog/f/data-visualization-using-the-fourier-transform.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[5] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[6] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020.

[7] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[8] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Plotly Technologies Inc. Collaborative data science, 2015.

[11] Luís Jesus, Ana Inês Tavares, and Andreia Hall. Cross-cultural adaption of the grbas and cape-v scales for portugal and a new training programme for perceptual voice evaluation. In *Advances in Speech-language Pathology*, pages 29–255. InTech, September 2017.

[12] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[13] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[14] Anand Mohan and Narayanankutty Karuppath. Real-time whispered to natural mode conversion of audio signals. *Available at SSRN 4778049*, 2023.

[15] E. Neuburg. Improvement of voicing decisions by use of context. In *ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3. IEEE, 1978.

[16] Gonçalo Duarte Nunes. Whispered speech segmentation based on deep learning. Master's thesis, University of Porto, 2023.

[17] Marco A. Oliveira. Machine learning approaches for whisper to normal speech conversion: A survey. *U. Porto Journal of Engineering*, 8(2):202–212, 2022.

[18] Daniel Pape and Luis M. T. Jesus. Stop and fricative devoicing in european portuguese, italian and german. *Language and Speech*, 58(2):224–246, June 2015.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] Olivier Perrotin and Ian V McLoughlin. Glottal flow synthesis for whisper-to-speech conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing; 28; 889–900*, 2020.

[21] R Pradeep, M Kiran Reddy, and K Sreenivasa Rao. Lstm-based robust voicing decision applied to dnn-based speech synthesis. *Automatic Control and Computer Sciences*, 53:328–332, 2019.

[22] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[23] M.W. Spratling. A review of predictive coding algorithms. *Brain and cognition*, (112):p. 92–97, 2017.

[24] Viet-Anh Tran and et al. Predicting f0 and voicing from nam-captured whispered speech. In *Speech Prosody 2008 - 4th International Conference on Speech Prosody*, 2008.

[25] Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

[26] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

Appendix A

SDG in Master's Dissertation

This appendix describes the possible impacts (albeit indirect in some regards) of this dissertation from the point of view of the United Nation's Sustainable Development Goal.

SDG	**Related Goal**	**Contribution**	**Effort Indicator**
SDG 3: Good Health and Well-being	Target 3.8: Achieve universal health coverage, including access to quality essential healthcare services	Investment in developing speech conversion technology directly contributes to advancing health technology	Research and Development Investment
		Developing and testing prototypes ensures that new healthcare solutions are safe and effective	Prototype Development and Testing
		Gathering feedback from users helps improve the technology to better meet health needs	User Engagement and Feedback
SDG 4: Quality Education	Target 4.5: Eliminate gender disparities in education and ensure equal access to all levels of education for the vulnerable	Academic dissemination supports education and awareness about speech impairments and solutions	Number of Research Papers and Publications
		Training programs for educators and healthcare providers ensure that they are equipped to support students with speech impairments	Training and Capacity Building

SDG	**Related Goal**	**Contribution**	**Effort Indicator**
SDG 9: Industry, Innovation, and Infrastructure	Target 9.5: Enhance scientific research, upgrade the technological capabilities of industrial sectors	Partnerships drive innovation and infrastructure development in speech technology	Collaborations and Partnerships
		Engaging skilled professionals in the project supports industrial innovation	Human Resources
SDG 10: Reduced Inequalities	Target 10.2: Empower and promote the social, economic, and political inclusion of all	Efforts to integrate the technology into healthcare and education systems promote social inclusion and reduce inequalities for individuals with speech impairments	Policy and Advocacy Efforts
		Widespread use of the technology helps ensure equal opportunities for individuals with speech difficulties	Implementation and Usage Rates
SDG 17: Partnerships for the Goals	Target 17.16: Enhance the global partnership for sustainable development	Cross-sector partnerships are essential for developing and implementing the technology, fostering collaboration for sustainable development	Collaborations and Partnerships
	Target 17.17: Encourage and promote effective public, public-private, and civil society partnerships	Engaging with policymakers ensures the integration of the technology in broader development agendas	Policy and Advocacy Efforts