UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA
CURSO DE GRADUAÇÃO EM ENGENHARIA COMPUTACIONAL

Caio Cedrola Rocha

Creating a Dataset for Automatic Phonetic Transcription in Brazilian Portuguese

Juiz de Fora
2024

Caio Cedrola Rocha

Creating a Dataset for Automatic Phonetic Transcription in Brazilian Portuguese

FULL TEXT IN ENGLISH.
Trabalho de conclusão de curso apresentado à Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de bacharel em Engenharia Computacional.

Orientador: Professor Doutor Jairo Francisco de Souza

Juiz de Fora

2024

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF com os dados fornecidos pelo(a) autor(a)

Rocha, Caio Cedrola.
Creating a Dataset for Automatic Phonetic Transcription in Brazilian Portuguese / Caio Cedrola Rocha. – 2024.
60 f. : il.

Orientador: Jairo Francisco de Souza
Trabalho de Conclusão de Curso – Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Curso de Graduação em Engenharia Computacional, 2024.

1. automatic phonetic transcription. 2. automatic speech recognition. 3. speech dataset. 4. brazilian portuguese. 5. grapheme-to-phoneme. I. de Souza, Jairo, orient. II. Título.

Caio Cedrola Rocha

Creating a Dataset for Automatic Phonetic Transcription in Brazilian Portuguese

FULL TEXT IN ENGLISH.
Trabalho de conclusão de curso apresentado à Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de bacharel em Engenharia Computacional.

Aprovada em 26 de junho de 2024

BANCA EXAMINADORA

Professor Doutor Jairo Francisco de Souza - Orientador
Universidade Federal de Juiz de Fora

Professor Doutor Heder Soares Bernardino
Universidade Federal de Juiz de Fora

Mestre José Eduardo de Carvalho Silva
Fundação CAEd

Dedico este trabalho a todos os meus professores da escola, que, embora indiretamente, tiveram uma contribuição fundamental nesta pesquisa, desde a leitura e a escrita até a formulação do "porquê" deste trabalho e a compreensão da importância que a educação teve na minha formação.

AGRADECIMENTOS

Agradeço ao meu orientador, Jairo, por aceitar conduzir o meu trabalho de conclusão de curso, e por ser um exemplo de professor e de líder de pesquisa, me ensinando a importância da organização e do método de pesquisa.

Ao José Eduardo pela sua habilidade de compartilhar conhecimentos técnicos da forma mais intuitiva, que me agregaram substancialmente durante meu período de estágio e de fim da graduação.

A todos os profissionais do curso de Engenharia Computacional na Universidade Federal de Juiz de Fora pela dedicação de cada um na consolidação e no reconhecimento do curso.

Aos meus pais e às minhas irmãs pelo apoio inequiparável durante todo o período da graduação, e por sempre me motivarem a sonhar alto e a fazer o meu melhor.

Por fim, aos meus amigos de curso pela imensa dedicação em trabalhos, projetos, e estágio, e por terem sido igualmente importantes nesta jornada, tanto acadêmica, quanto pessoal.

"Minha pátria é a língua portuguesa." (PESSOA, 1982, p. 358).

RESUMO

A Transcrição Fonética Automática (APT) é a tecnologia que automatiza o processo de converter fala em transcrições fonéticas. Ela é crucial para melhorar a precisão dos sistemas de Reconhecimento Automático de Fala (ASR). Modelos de aprendizado profundo, como wav2vec 2.0, têm mostrado desempenho notável em aprender características fonéticas a partir de dados. No entanto, eles requerem corpora de fala transcritos foneticamente, que são escassos em idiomas como o Português Brasileiro (PT-BR). O principal objetivo desta pesquisa é estabelecer uma abordagem sistemática para gerar um conjunto de dados com transcrições fonéticas automáticas para PT-BR a partir de corpora de ASR disponíveis. Utilizando ferramentas de conversão de Grafema para Fonema (G2P), o objetivo é otimizar o processo de transcrição e aprimorar o treinamento dos modelos APT. Pesquisamos corpora de fala em PT-BR adequados para treinar modelos APT, selecionando, por fim, o corpus CORAA ASR. Além disso, avaliamos cinco conversores G2P para PT-BR, padronizando as transcrições segundo um quadro referência de fonemas em PT-BR. O conversor G2P do FalaBrasil alcançou a menor taxa de discordância entre as ferramentas selecionadas, e foi usado para transcrever o corpus CORAA ASR utilizado para o ajuste do modelo fonético. O ajuste fino em 10 horas de áudio retornou uma taxa de erro de fonemas (PER) de 15,87% no conjunto de testes. Outrossim, o modelo apresentou altas pontuações médias de confiança por fonema, bem como pouca confusão entre fonemas, e foi compartilhado no repositório da Hugging Face, contribuindo para a pesquisa de ASR em PT-BR.

Palavras-chave: transcrição fonética automática, reconhecimento automático de fala, conjunto de dados de fala, português brasileiro, grafema para fonema

ABSTRACT

Automatic Phonetic Transcription (APT) is the technology that automates the process of converting speech into phonetic transcriptions. It is crucial for improving the accuracy of Automatic Speech Recognition (ASR) systems. Deep learning frameworks, such as wav2vec 2.0, have shown remarkable performance in learning phonetic features from data. However, they require phonetically transcribed speech corpora, which are scarse in languages such as Brazilian Portuguese (PT-BR). The primary objective of this research is to establish a systematic approach for generating a dataset with automatic phonetic transcriptions for PT-BR from available ASR corpora. By leveraging Grapheme-to-Phoneme (G2P) conversion tools, the aim is to streamline the transcription process and enhance the training of APT models. We researched PT-BR speech corpora suitable for training APT models, ultimately selecting the CORAA ASR corpus. Additionally, we evaluated five G2P converters for PT-BR, standardizing the transcriptions according to a reference phoneme chart. FalaBrasil's G2P achieved the lowest discordance rate among the selected G2P tools, leading to its selection for transcribing the CORAA ASR corpus used in fine-tuning. The fine-tuning on 10 hours of audio yielded a 15.87% PER (Phoneme Error Rate) on the test set. In addition, the model presented high average confidence scores per phoneme, as well as little confusion between phonemes. It was then shared on the Hugging Face repository, contributing to ASR research in PT-BR.

Keywords: automatic phonetic transcription, automatic speech recognition, speech dataset, brazilian portuguese, grapheme-to-phoneme

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ASR	Automatic Speech Recognition
DNN	Deep Neural Network
APT	Automatic Phonetic Transcription
G2P	Grapheme-to-Phoneme
PT-BR	Brazilian Portuguese
PT-PT	European Portuguese
MLP	Multilayer Perceptron
RNN	Recursive Neural Network
APR	Automatic Phoneme Recognition
CTC	Connectionist Temporal Classification
WER	Word Error Rate
CER	Character Error Rate
PER	Phoneme Error Rate
IPA	International Phonetic Alphabet
X-SAMPA	Extended Speech Assessment Methods Phonetic Alphabet
TTS	Text-to-Speech
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficient
CNN	Convolutional Neural Network
PASE	Problem Agnostic Speech Encoder

LIST OF SYMBOLS

$f : A \to B$	Function with domain A and codomain B

CONTENTS

1 Introduction

Automatic speech recognition (ASR) is an important technology to enable and improve the human–human and human–computer interactions. Speech technology has changed the way we live and work and has become one of the primary means for humans to interact with some devices (DENG, 2016).

ASR as user interface has become ever more useful and pervasive (LI et al., 2015), embodied in voice search, short message dictation, and virtual speech assistants (DENG, 2016). ASR can also support the medical and education sectors (ALHARBI et al., 2021). Equally important is the development of deep learning techniques powered by big data and significantly increased computing ability (DENG, 2016).

As highlighted by Nassif et al. (2019), the majority of papers published since 2006 have employed deep neural networks (DNNs) for ASR. They benefit from the large amount of speech data and transcriptions available on the internet. To train such networks, a large dataset of audio data is essential to ensure robustness and generalization in the transcriptions.

Automatic Phonetic Transcription (APT) is subfield of ASR that focuses on transcribing speech in the phoneme level. APT can be applied in fields such as phonology, linguistics, education, and medicine, where detailed phonetic analysis is required. In this work, we center our attention on the critical role of a labeled dataset for improving the accuracy of APT models, emphasizing how such a resource is instrumental in advancing the effectiveness of phonetic transcription tasks.

1.1 Problem description

The industry has developed a broad range of commercial products where ASR as a consumer-centric technology increasingly require robustness in noisy everyday environments. However, reliably recognizing spoken words in realistic acoustic environments is still a challenge (LI et al., 2015). In this sense, accurate phoneme recognition is crucial to the improvement of ASR systems. Bhatt et al. (2023) stated that phoneme recognition is influenced by the duration of the phoneme, speaking rate, style, accent, contextual effects, age, gender, health condition, training and testing environments. The study also exposed that existing works suffered with the confusion of phonemes within the same category and the lack of state-of-the-art resources.

In addition, the training of APT models with manual phonetic transcriptions is costly and impractical due to the lack of such resource for most speech recognition tasks. According to Van Bael et al. (2007), manual verification of phonetic transcriptions is time-consuming and expensive. Several challenges revolve around obtaining phonetic transcriptions - the time required, the high costs incurred, the often limited accuracy

obtained, and, especially for speech technology applications, the need to transcribe large amounts of data. It is equally important to consider the subjectivity of human transcriptions and the complex methodology needed to generate them (CUCCHIARINI; STRIK, 2003). Especially for under-resourced languages, ensuring reliability and validity can be challenging due to the lack of trained professionals available to engage in this task.

An effective approach for generating phonetic transcriptions to train APT systems is to use Grapheme-to-Phoneme (G2P) conversion. G2P convert orthographic transcriptions, commonly available in speech corpora, into phonetic transcriptions. This process is especially useful for Portuguese, where datasets with phonetic transcriptions are limited, and G2P tools targeting this language can quickly produce such dataset from other available datasets. Even so, the absence of comprehensive studies evaluating the accuracy of G2P converters for ASR in Portuguese raises questions about whether these converters meet the expected standards. There is no consensus on which phoneme representation format, G2P algorithm, or lexicon symbolizes the state of the art in Portuguese, which further underscores the complexity of this issue (SANTOS; BARONE; ADAMI, 2008). There are also other factors involved, such as the different accents within Portuguese, that need each their own phonetic representations.

1.2 Justification

Bhatt et al. (2023) argued that there is a need to investigate phoneme recognition to improve speech recognition as an accurate recognition of phonemes leads to improved speech recognition. Van Bael et al. (2007) investigated several studies that reported the benefits of using APTs for the development of ASR systems. APT can provide more nuanced information about the spoken units and aid in the recognition of words. Therefore, the determination of a vocabulary of phonemes and a dataset of phonetic transcriptions suitable for training APT models for Portuguese serves as a foundation to which researchers and students can build upon. Ideally, this research also aims to provide a standard that can be adapted to other languages as well, especially since several G2P converters are already readily accessible for a wide range of languages.

Cucchiarini e Strik (2003) claims that APTs offer several advantages for phonetic research. One advantage is that they can achieve uniformity in phonetic transcription, which is difficult to achieve with human transcribers. Another advantage is that they can generate phonetic transcriptions of large amounts of data that would otherwise be too time-consuming and expensive to transcribe manually. Furthermore, APTs can help control for biases that may be present in human transcriptions.

The possibility of transcribing readily available speech corpora using G2P holds the promise of streamlining the training process for APT models, producing fairly reliable training phonetic transcriptions. In addition, by providing a phonetic transcription

dataset, this work could contribute to enhancing the versatility of APT models, potentially benefiting applications like voice assistants, phonetic transcription tasks, and more.

Lastly, by sharing our fine-tuned APT model for Brazilian Portuguese (PT-BR) using the latest ASR technology, wav2vec 2.0, we aim to offer an updated perspective on APT models for this language. Our model will be ready for out-of-the-box APT tasks.

1.3 Objectives

The main objectives of this research revolve around addressing challenges in APT for PT-BR. Firstly, the focus is on developing a robust methodology to create a labeled dataset for phonetic transcription. This was done by using G2P tools to produce phonetic transcriptions from existing graphemic annotations of speech corpora. The aim is to provide accurate phonetic transcriptions to facilitate the training of APT models, overcoming the limitations associated with manual transcriptions. This approach not only streamlines the training process but also enhances the scalability and accessibility of APT models.

Secondly, our objective is to enrich the resources available for the utilization and fine-tuning of APT models in the PT-BR language. This entails sharing a PT-BR APT model fine-tuned on the proposed dataset, accompanied by performance metrics and a comparative analysis with existing work. Through this initiative, we seek to contribute not only to the accessibility of specialized APT models but also to the benchmarking and further refinement of APT technology for PT-BR.

1.4 Organization

This document is organized into 5 chapters. Chapter 2 provides the theoretical foundation necessary to fully understand the conducted work. The topics addressed are ASR, speech corpora, phonetic representations, and G2P conversion. The chapter also presents a selection of related work and an analysis of similarities and differences between them. Chapter 3 outlines the materials and methods used in the development of our work. Chapter 4 presents the results and an analysis of the phonetic transcriptions and the APT models. Finally, Chapter 5 discusses the conclusions and suggests areas for future research.

2 Literature Review

This chapter presents a critical exploration of existing works relevant to the field of APT and its applications in PT-BR. We aim to provide a comprehensive overview of the current state of knowledge, seeking to contextualize APT in ASR and highlight key concepts, common practices, and state-of-the-art methods. This examination is integral to shaping the theoretical framework and methodology of the current research, contributing to a deeper understanding of the proposed dataset.

2.1 Theoretical Foundation

This section establishes the conceptual framework for our study on APT in PT-BR, serving as a guide to the subsequent analysis and methodology to contribute to the understanding of APT in the context of PT-BR.

2.1.1 ASR

According to Li et al. (2015), "ASR is the process and the related technology for converting the speech signal into its corresponding sequence of words or other linguistic entities by means of algorithms implemented in a device, a computer, or computer clusters". In other words, ASR is a technology that aims to convert human speech into syntactic and semantic features of language, enabling machines to comprehend and interpret spoken language. Essentially, it involves capturing a relevant signal and transforming it into pertinent information, i.e., recognizing a pattern in the (speech) signal (O'SHAUGHNESSY, 2008).

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words, independent of the device used to record the speech (i.e. microphone), the speaker's accent, or the acoustic environment in which the speaker is located (e.g., quiet office, noisy room, outdoors) (RABINER; SCHAFER, 2007). ASR tasks vary in 4 dimensions, presented below (JURAFSKY; MARTIN, 2023):

- Vocabulary size. The vocabulary can contain up to 60,000 words in open-ended tasks like transcribing videos or human conversations.
- Context. Read the speech, in which humans read aloud, for example in audiobooks, is easier to recognize than conversational speech, for example, for transcribing a business meeting.
- Channel and noise. Speech is easier to recognize if it's recorded in a quiet room with head-mounted microphones than if it's recorded by a distant microphone on a

noisy city street, or in a car with the window open.

- Accent or speaker-class characteristics. Speech is easier to recognize if the speaker is speaking the same dialect or variety that the system was trained on, while speech by speakers of regional or ethnic dialects, or speech by children can be quite difficult to recognize if the training data is not diverse enough.

On the other hand, APT concentrates on the transcription of phonemes in speech. It consists of finding the best possible sequence of phonemes that are contained in a given speech sample (LOPES; PERDIGAO, 2011). The classification and recognition of phonemes are considered the primary tasks of ASR systems irrespective of application domain (MALAKAR; KESKAR, 2021).

2.1.1.1 Deep learning for APT

Deep learning architectures have revolutionized the field of speech processing by demonstrating remarkable performance across various tasks (MEHRISH et al., 2023). DNNs are conventional multilayer perceptrons (MLPs) with many hidden layers and are capable of learning from unstructured data by using fine-tuning done with the backpropagation technique (KUNAPULI; BHALLAMUDI, 2021). DNNs automate the feature extraction processes and observe patterns in the data that allow them to cluster inputs appropriately. It has been shown that they can capture discriminative information regarding the internal structure of phonetic data (RIZWAN; ANDERSON, 2016).

Traditional DNN approaches like Recursive Neural Networks (RNNs) have been successfully applied to Automatic Phoneme Recognition (APR), including phone probability estimation, phoneme classification, sequence labeling, phoneme recognition, and acoustic modeling (MALAKAR; KESKAR, 2021). However, as supervised learning models, RNNs require the building of specialist models for individual tasks and application scenarios that require a significant amount of labeled training data. Thus, it is difficult to apply this to dialects and languages for which there are limited resources (MOHAMED et al., 2022).

In this context, Transformer-based approaches have gained more attention for their ability to perform exceptionally well in end-to-end speech recognition systems (YEH et al., 2019). Transformers learn the representations of speech via self-supervised learning, which involves unsupervised pre-training followed by supervised fine-tuning. Self-supervised learning can easily incorporate expert-derived priors into the training process by tasking the model to recover known signal transformations that are repetitively derived without the need for annotated data (RAVANELLI et al., 2020). This is exceptionally convenient in scenarios where there is a limited amount of annotated training data, as is often the case in APT. In this work, we concentrate on leveraging wav2vec 2.0, a framework based on

a Transformer model, for phonetic transcription in Portuguese. Additionally, we explore multilingual approaches, which offer the flexibility to recognize phonemes associated with the Portuguese language.

2.1.1.2 wav2vec 2.0

wav2vec 2.0 is a framework for self-supervised learning of representations from raw audio data. It is able to learn powerful representations from unlabeled speech data (BAEVSKI et al., 2020), including speech units common to several languages. This capability proves beneficial in scenarios where there are small amounts of unlabeled speech since under-resourced languages can benefit from languages for which more data is available (BAEVSKI; CONNEAU; AULI, 2020). The process of learning from large amounts of unannotated data is known as pre-training, which allows the model to adjust parameters and learn the underlying speech units.

Mihajlik, Révész e Tatai (2002) cite that a key point in adjusting the parameters of a speech recognizer is the need for the transcription of the recorded training speech. The same holds true for the pre-trained wav2vec 2.0 models, which need to be fine-tuned for a specific task, though with a much smaller amount of annotated data, to perform speech-to-text transcription. In the case of APT, having an annotated dataset of phonetic transcriptions is of paramount importance to the model fine-tuning, because it allows the model to recognize a specific set of phonemes from the audio signal and carry out the phonetic transcription.

The wav2vec 2.0 model architecture is displayed in Figure 1.

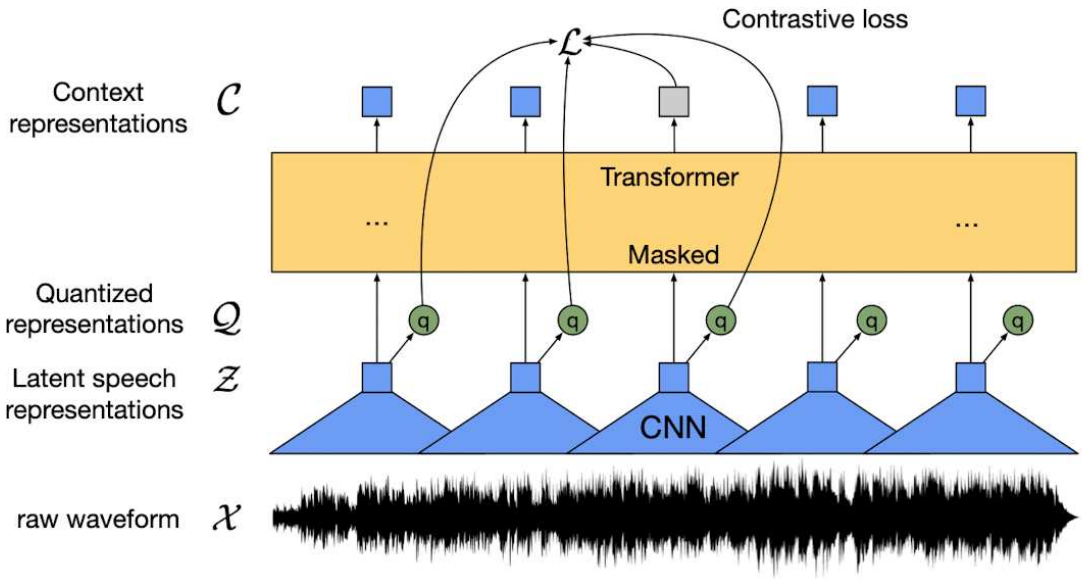


Figure 1 – wav2vec 2.0 architecture. Source: (BAEVSKI et al., 2020).

The model is composed of a multi-layer convolutional feature encoder $f : X \rightarrow Z$

which takes as input raw audio X and outputs latent speech representations $z_1, ..., z_T$ for T time-steps. They are then fed to a Transformer $g : Z \rightarrow C$ to build contextualized representations $c_1, ..., c_T$ capturing information from the entire sequence. The output of the feature encoder is discretized to q_t with a quantization module $Z \rightarrow Q$ to represent the targets in the self-supervised objective. The benefit of wav2vec 2.0 is that it builds context representations over continuous speech features, and self-attention captures dependencies over the entire sequence of latent representations end-to-end.

To pre-train the model, a certain portion of Z is masked, the aim of the pre-training step is not to reconstruct the continuous representations directly, but to identify the correct quantized latent representation for the masked time step among a set of distractors. The representations of speech are learned by solving a contrastive task $\mathcal{L}_m$, which requires distinguishing the true quantized latent speech representation from the distractors. This is augmented by a diversity loss $\mathcal{L}_d$ to encourage the model to use possible speech units equally often. The total loss is:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_d$$

After pre-training on unlabeled speech, the model can be fine-tuned on labeled data with a Connectionist Temporal Classification (CTC) objective to be used for downstream speech recognition tasks (BAEVSKI et al., 2020).

2.1.1.3 Evaluation Metrics

The information within this subsection is predominantly sourced from the audio course authored by Gandhi et al. (2023).

When assessing speech recognition systems, we compare the system's predictions to the target transcriptions. There are three categories of errors, which can be computed on the word level or on the character level:

- Substitutions (S): where the wrong word (or character) is transcribed ("sit" instead of "sat")
- Insertions (I): where there is an extra word (or character) in the transcription
- Deletions (D): where a word was removed from the transcription

When assessing the performance of an ASR model, the Word Error Rate (WER) is the *de facto* metric. It calculates substitutions, insertions, and deletions on the word level, according to equation 2.1.

$$WER = \frac{S + I + D}{N} \tag{2.1}$$

In Equation 2.1, N represents the number of words in the reference transcription. A lower WER signifies fewer errors in the system's transcriptions. The ideal speech recognition system would achieve a WER of zero.

In turn, the Character Error Rate (CER) assesses systems on the character level, according to equation 2.2.

$$CER = \frac{S + I + D}{N} \tag{2.2}$$

In Equation 2.2 N represents the number of characters in the reference transcription. Likewise, a lower CER indicates better performance. The Phone Error Rate (PER) is an alternative to the CER commonly used in the evaluation of APT models. It is calculated as the CER, except that each phoneme is considered an individual character.

For ASR systems, WER is usually favored over CER due to its emphasis on contextual understanding, requiring systems to accurately grasp the context of predictions. For instance, if the system predicts "sit" instead of the correct tense "sat", it indicates a failure to comprehend the relationship between verb and tense in the sentence. Therefore, it provides a more comprehensive evaluation of system performance. However, in specific applications like phoneme recognition systems, CER becomes preferable because it assesses precision at a finer, character-level granularity.

A common metric in APT utilized to identify the accuracy of the predicted phonemes is the confusion matrix, which represents the prediction summary in matrix form, showing how many predictions are correct and incorrect per class. It helps in understanding the classes that are being confused by the model as other classes (TIWARI, 2022). In its normalized form, the confusion matrix is adjusted to express values within the 0-1 range. This normalization is achieved by scaling the rows to the 0-1 range, providing a more intuitive representation of prediction accuracies. As an example, for the transcription /jɛə/ ("yeah") predicted as /jeə/, we would have the confusion matrix given in Table 1.

		Reference			
		e	**j**	**ə**	**ɛ**
Predicted	**e**	0	0	0	1
	j	0	1	0	0
	ə	0	0	1	0
	ɛ	0	0	0	0

Table 1 – Example confusion matrix.

Table 1 compares the phonemes predicted by the system against the actual or reference phonemes from the dataset. The matrix is organized into rows and columns, where each row corresponds to the reference phonemes, and each column corresponds to the predicted phonemes. The four main components of a confusion matrix are:

- True Positives (TP): The number of phonemes correctly predicted by the system, where both the reference and predicted phonemes match.
- True Negatives (TN): The number of instances where non-occurring phonemes are correctly identified as such.
- False Positives (FP): The number of times the system predicts a phoneme that is not present in the reference, indicating an incorrect positive prediction.
- False Negatives (FN): The number of times the system fails to predict a phoneme that is present in the reference, indicating a missed positive prediction.

In addition, the accuracy represents the overall correctness of the phoneme recognition system. It is calculated as the ratio of the sum of true positives (correctly predicted phonemes) and true negatives (correctly identified non-occurring phonemes) to the total number of predictions. Equation 2.3 displays the formula for accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

The accuracy can also be calculated as the ratio of the sum of the main diagonal of the confusion matrix and the sum of the confusion matrix. In simpler terms, accuracy measures the proportion of correctly identified phonemes (both positive and negative) out of all the predictions made by the system. A high accuracy value indicates that the system is performing well in correctly identifying phonemes, while a lower accuracy suggests more errors in the predictions. The accuracy of the example above would be:

$$Accuracy = \frac{2}{3} \approx 0.66$$

2.1.2 Speech Corpora

A speech corpus (plural speech corpora) is a large collection of audio recordings of spoken language, optionally accompanied by additional text files containing transcriptions of the words spoken and the time each word occurred in the recording. Speech corpora can be divided into two types: read speech, also known as prepared speech, which includes excerpts from books, news broadcasts, word lists, and number sequences; and spontaneous speech, which includes narratives, meetings, debates, dialogs, and phone conversations (RICHEY, 2020).

There are several corpora available to ASR destined for different purposes, each with different types of annotation. According to Raso e Mello (2012), “the value of a corpus can only be measured in terms of its success in meeting the purposes for which it was created”. Therefore, it is foremost to understand the context of application for

which a corpus was created and the type of annotation used. In order for it to be useful for research it needs to be labeled in some way, with a minimum requirement being the transcription of spoken words in standard orthography. Sometimes additional linguistic information can be provided: syllables, sounds, intonation, disfluencies, filled pauses (e.g. “um”, “uh”), and the phonetic transcription (RICHEY, 2020).

Some speech corpora are provided with a phonemic lexicon that can be used to generate a hypothetical canonical phonetic representation of the orthographic transcripts. In addition, a handful of speech corpora may partially provide broad phonetic transcriptions with the help of human transcribers in order to ensure a more accurate representation of the material (Van Bael et al., 2007).

2.1.3 Phonetic Representations

The preparation of a dataset of phonetic transcriptions require the adoption of notational standards for representing these transcriptions, as well as, criteria for deciding which phonemes the dataset will encompass. The following subsections address the key elements utilized to take such decisions.

2.1.3.1 International Phonetic Alphabet

The International Phonetic Alphabet (IPA) is an international alphabet used by linguists to accurately represent the wide variety of sounds (phones or phonemes) in human speech, aiming to provide the academic community worldwide with a notational standard for the phonetic representation of all languages. The IPA is an important resource for allowing accurate phonetic transcriptions in a language, and it will be used from here on to represent the phonemes in PT-BR.

2.1.3.2 PT-BR Phonemes

There is no consensus about which phonemes are present in PT-BR, therefore there is no unified phoneme chart. For this work, we adopt the set of consonants present in Ivo (2019a) (Figure 2) and the set of vowels present in Ivo (2019b) (Figure 3).

Figure 2 displays similar sounds side by side in each cell, which account for sounds that could be considered equivalent in Portuguese. In addition to the phonemes displayed in Figure 2, we also opted to include the consonants /w/, /w̃/, /j/, and /j̃/ because they appeared in the examples of the cited material.

We stress the difference between phones and phonemes: the first are the physical speech sounds, that is the language-independent units shared by all languages (LI et al., 2021). In contrast, phonemes are the minimal — language-dependent — units that distinguish between words, i.e., between one meaning and another (MITTERER; CUTLER, 2006). Allophones are sets of phones within a language that correspond to the same

Manner \ Place	Bilabial	Labio-dental	Dental or Alveolar	Alveopalatal	Palatal	Velar	Glottal
Stop	p b		t d			k g	
Affricate				t͡ʃ d͡ʒ			
Fricative		f v	s z	ʃ ʒ		x ɣ	h ɦ
Nasal	m		n		ɲ ỹ		
Tap			ɾ				
Trill			r				
Retroflex			ɹ				
Lateral			l ɬ		ʎ ʟ		

Figure 2 – Phoneme chart of PT-BR consonants

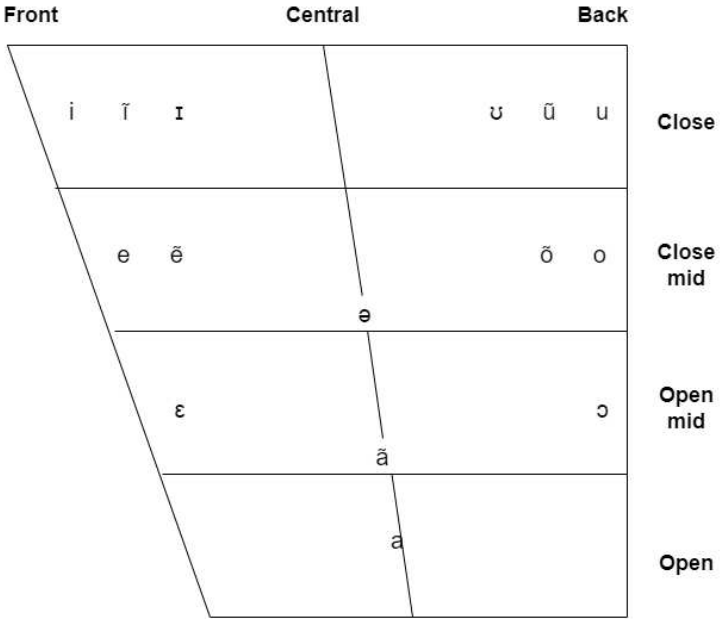


Figure 3 – Phoneme chart of PT-BR vowels

phoneme; although indistinguishable within that language, they may constitute separate phonemes in other languages (LI et al., 2020).

The so-called Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) is a keyboard-compatible coding for the entire set of IPA symbols that aims to offer the basis of an international standard machine-readable phonetic alphabet (WELLS, 1995). The mapping of X-SAMPA and IPA symbols used in this work is presented in Appendix A.

2.1.4 G2P converters

These tools are typically language-specific and hold significant importance in the context of APR due to their capability to generate labeled datasets for training APT models. A study by Jouvet, Fohr e Illina (2012) shows that the training process of ASR systems that use transcriptions derived from G2P models is quite robust to some errors in the pronunciation lexicon, whereas pronunciation lexicon errors are harmful in the decoding process of the neural network. Thus, using a robust G2P converter is foremost for minimizing errors in the APT model.

G2P models convert words to their phonetic pronunciations. They are an important module in text-to-speech (TTS) and ASR systems. Two main G2P approaches are: knowledge-based (rule-based) and data-driven (SAR; TAN, 2019). Knowledge-based models use pre-defined rules and pronunciation dictionaries to generate phonetic transcriptions but they are highly language-dependent and require expert linguistic knowledge to design. Hence they require significant manual effort to build, and have limited adaptivity on unseen words. On the other hand, data-driven methods incorporate learning, such as Long Short-Term Memory (LSTM) and Transformer-based attention models (ENGELHART; ELYASI; BHARAJ, 2021). They learn from a corpus of word-phonetic transcription pairs without explicit modeling of linguistic knowledge. The advantages of such an approach are that the technique is reusable for different sets of data (e.g. different languages or dialects) and are more robust against unseen words (BOSCH; DAELEMANS, 1993).

2.2 Related Work

In this section, we survey the existing literature on APR, with a specific emphasis on the training data, G2P methods, and APT strategies. The review not only explores the state-of-the-art methods but also highlights similarities and differences among select works, offering comparisons with the proposed methodology — specifically, the preparation of a training dataset for APT in PT-BR.

The identification of relevant literature was facilitated by the Google Scholar search engine, employing keywords such as “Automatic Speech Recognition”, “phonetic transcription”, “automatic phonetic transcription”, “automatic phoneme recognition”, “Portuguese”, “Brazilian Portuguese”, “multilingual”, and “dataset”. Criteria for selection included prioritizing recent publications, similarity in methodology, alignment of goals, and a focus on either multilingual APT or APT specific to PT-BR.

2.2.1 Datasets

Aguiar e Costa-Abreu (2023) proposed a methodology to collect and release a speech dataset for PT-BR. The data was collected from a playlist of TEDx talks in

Portuguese available on the TEDx channel. The authors used available automatic or human-generated captions and performed a forced alignment with the Montreal Forced Aligner to generate the time intervals for each word and phoneme in the captions. This approach proved very useful because the authors could release a large PT-BR speech dataset with phonetic annotations, though it is not yet publicly available. The dataset also features a variety of Portuguese accents and demographic information. This study relates to the current thesis for providing the methodology used to generate the dataset, though it creates a new speech corpus, instead of utilizing existing ones.

Dijkstra (2021) employed a comparable approach when it comes to the phonetic transcription. They used Praat with the EasyAlign plugin to transcribe the Sid and LaPS Benchmark 16k PT-BR datasets, obtaining phonetic transcriptions and alignments. While this study is closely aligned with the current research in providing phonetic transcripts for PT-BR corpora, our work proposes the use of G2P converters for faster and more adaptable transcription. However, this study stands out for offering phoneme alignments in PT-BR, valuable information for various ASR systems.

Two of the selected works targeted specifically APT for low-resource languages. Li et al. (2020) focused on phone recognition for two indigenous languages. They derived phonemic transcriptions from standard transcriptions of 12 rich-resource languages using the Epitran G2P tool Mortensen, Dalmia e Littell (2018) and trained a multilingual allophone system. The study proposed a mapping of articulatory features, thus being able to infer phonetic information for unknown phonemes, which is very useful for multilingual APT. Using a slightly different approach, Yi et al. (2021) fine-tuned wav2vec 2.0 models on the CALLHOME corpus for six low-resource languages, mapping graphemes to phonemes through lexicons provided by the corpus.

Xu, Baevski e Auli (2022) trained a single multilingual model on labeled data from several languages, enabling zero-shot phoneme recognition. The training data covered 26 languages from the Common Voice corpus, including PT-BR; 19 languages from the Babel corpus; and six languages from the Multilingual LibriSpeech corpus. They used Espeak and Phonetisaurus G2P converters for obtaining phonetic transcriptions. Similarly, Taguchi et al. (2023) used training data from seven languages of CommonVoice 11.0, transcribing them into IPA semi-automatically, and employed both Epitran and manual rules for G2P conversion, manually checking the quality of phonetic transcriptions. In addition, the study assembled a test set of manually annotated phonetic transcriptions of four low-resource languages from Common Voice.

Klumpp et al. (2022) introduced Common Phone, a gender-balanced, multilingual corpus with approximately 116 hours of speech from six languages of the Common Voice corpus. The corpus was enriched with automatically generated phonetic segmentation. This was done by estimating the true pronunciation from the ideal (G2P) and recognized

(ASR) pronunciations. The training split from Common Phone was used to fine-tune a wav2vec 2.0 base model. While providing a valuable speech corpus for APT, it is important to note that this corpus does not include PT-BR audio recordings.

Quintanilha (2017) developed a dataset from an ensemble of four PT-BR speech datasets. While not specifically designed for APT, this work offers a detailed overview of ASR, end-to-end speech recognition, and preparation of the training, test, and evaluation datasets, utilizing popular PT-BR speech corpora. Junior et al. (2023) presented CORAA (Corpus of Annotated Audios) ASR, a publicly available dataset for PT-BR ASR, featuring 290 hours of audios with validated transcriptions. The corpus includes both prepared and spontaneous speech, and was assembled with the goal of improving PT-BR ASR models.

In short, for the related works that targeted APT, Li et al. (2020), Taguchi et al. (2023), Yi et al. (2021), and Xu, Baevski e Auli (2022) employed G2P methods to generate the phonetic transcriptions; and Aguiar e Costa-Abreu (2023), Klumpp et al. (2022) and Dijkstra (2021) relied on software to infer the phonetic transcriptions from audio inputs and their corresponding orthographic transcriptions.

2.2.2 APT strategies

Dijkstra (2021) developed a technique for APT targeting PT-BR that uses Mel Frequency Cepstral Coefficients (MFCC) and filter banks for acoustic processing and convolutional neural networks (CNNs) together with LSTMs for phoneme classification. The Kaldi Speech Recognition Toolkit was used to extract the spectral features from the sound samples. Then, CNNs augmented with LSTMs were trained for phoneme classification with an output CTC layer to obtain the transcriptions. In addition to PT-BR, the authors also used the TIMIT dataset for English APT. In contrast to this work, our work aims to evaluate the performance of APT with wav2vec 2.0, which has been more widely used in the recent literature.

Nedjah, Bonilla e de Macedo Mourelle (2023) proposed an ASR system based on phoneme identification for PT-BR that can classify phones in real time through a set of spectral characteristics extracted from the speech signal. The authors used an ensemble of neural network experts, allowing to divide the decision space, and a RNN as a dynamic post-processing step to mitigate the oscillation generated by the static classification of samples. This work differs from the current thesis for performing a continuous classification of isolated phonemes, rather than a continuous phonetic transcription of words. Dijkstra (2021) mentions that the identification of speech units does not require the identification of pauses. In contrast, APT is a more complex task that requires recognizing words and syllables.

The task of multilingual APT and zero-shot phoneme recognition is addressed in

several studies. Xu, Baevski e Auli (2022), Klumpp et al. (2022), and Taguchi et al. (2023) presented the fine-tuning of multilingual wav2vec 2.0 models for speech-to-IPA transcriptions, taking advantage of previously learned phones in several languages to identify phonemes in other languages. Xu, Baevski e Auli (2022) stands out for proposing a mapping of phonemes of training languages to target languages using articulatory features, such as voicing or the place and manner of articulation. These studies share similarities with the current thesis, as they all employ multilingual wav2vec 2.0 models to infer phonemes in a target language. However, the proposed models have an extensive vocabulary, a characteristic that may introduce potential ambiguity and variability in the phonetic transcriptions. This thesis aims to present a simpler model with a shorter vocabulary.

Li et al. (2020) presented a self-supervised learning model for multilingual phoneme recognition that incorporates knowledge of phonology through an allophone layer. They also explicitly modeled language-independent phones, building a universal phone recognizer that, combined with a manually curated database of phone inventories, can be customized into 2,000 language dependent recognizers. This work showcases the potential of creating adaptable and language-specific phoneme recognition models from multilingual ASR systems. It differs from the previous papers for using a modified Problem Agnostic Speech Encoder (PASE+) instead of wav2vec 2.0. This architecture combines a convolutional encoder followed by multiple neural networks, called workers, tasked to solve self-supervised problems (LI et al., 2020).

Regarding studies focusing on ASR rather than APT for PT-BR, Quintanilha (2017) presented a character-based end-to-end speech recognition system for PT-BR using LSTM and CTC, varying the number of layers, applying different regularization methods, and fine-tuning several other hyperparameters. This study contributes to the understanding of training and fine-tuning ASR models. In contrast, Junior et al. (2023) introduced a fine-tuned version of the wav2vec 2.0 XLSR-53 model originally proposed by Ruder, Søgaard e Vulić (2019). This study serves as a potential benchmark for comparing WERs with other wav2vec 2.0 solutions in PT-BR, including APT systems. Similarly, Grosman (2021), Grosman (2022), and Gris et al. (2022) have shared fine-tuned wav2vec 2.0 models tailored for ASR in PT-BR.

Furthermore, AI at Meta provides a fine-tuned version for PT-BR of XLSR-53 (RUDER; SØGAARD; VULIĆ, 2019). These models are freely available on the Hugging Face repository. However, the absence of a fine-tuned wav2vec 2.0 APT model for Portuguese highlights the importance of sharing one specifically tailored for this language.

In summary, the reviewed literature offers a comprehensive overview of both multilingual and PT-BR APT resources. The studies discussed not only highlight the challenges associated with APT but also introduce innovative methodologies to address the

persistent issue of limited phonetic transcriptions. Notably, Dijkstra (2021) and Junior et al. (2023) have made significant contributions by introducing updated APT and ASR resources, respectively, in PT-BR. However, the existing gap in the availability of speech corpora with comprehensive phonemic transcriptions for PT-BR underscores the necessity for a more targeted approach. The current work addresses this gap by providing a dataset of phonetic transcriptions derived from existing PT-BR speech corpora. Additionally, sharing a fine-tuned APT model with this dataset aims to catalyze further advancements in the evolving field of APT for PT-BR.

3 Materials and Methods

This chapter outlines the procedures employed to address the challenges identified in the related work and contribute to the field of APT for PT-BR. The aim is to provide a clear and replicable framework for the development of a training dataset and the sharing of a fine-tuned APT model. To attain these objectives, we have outlined the following workflow:

1. Research and curate PT-BR speech corpora suitable for ASR, and assess their representativeness and suitability for robust APT model training;
2. Analyze the characteristics of the audios and of the speakers in the selected speech corpus;
3. Adopt a reference PT-BR phoneme chart;
4. Investigate and evaluate G2P converters for PT-BR, considering factors such as discordance, accent, and suitability for converting a large transcription dataset;
5. Adjust the G2P outputs to match reference phoneme chart and define the training vocabulary;
6. Fine-tune APT models on the training set and assess their PERs on the test set;
7. Analyze PERs across different phoneme groups, identifying phonemes with the highest and lowest error rates;
8. Document the process of creating the dataset and share it, together with the fine-tuned model;

3.1 Dataset

To compile a dataset of accurate phonetic transcriptions for PT-BR, extensive research of speech corpora for ASR in this language was conducted. This included an analysis of corpus types, context, objectives, size, and audio quality. Standard Google and Google Scholar search engines were employed, utilizing combinations of keywords such as "speech corpus", "Brazilian Portuguese", "multilingual", "Automatic Speech Recognition", and "phonetic transcription" in both English and Portuguese to identify relevant resources. The relevant information for the identified speech corpora is compiled in Table 2.

Name	Speech type	Size	Type of annotation	Number of speakers	Type of environment	License	Price	Last updated
CORAA ASR	Spontaneous and prepared	290h46min	Orthographic	∼1700	Both	CC BY-NC-ND 4.0	Free	2023
Common Voice	Read	204h	Orthographic	3247	Uncontrolled	CC-0	Free	2023
VoxForge	Read	4,130 utterances	Orthographic	180	Uncontrolled	GPL	Free	2023
MF: Male/Female for Forced Phonetic Alignment	Read	15min	Orthographic and phonetic	2			Free	2022
Multilingual TEDx	Prepared	164h	Orthographic			CC BY-NC-ND 4.0	Free	2021
Multilingual LibriSpeech	Read	284h35min	Orthographic	31		CC BY 4.0	Free	2020
CETUC	Read	145h	Orthographic	101	Controlled		Free	2019
LaPS Benchmark	Read	0h54min	Orthographic	35	Uncontrolled		Free	2019
Constituição	Read	7h11min	Orthographic	1	Controlled		Free	2018
Código de Defesa do Consumidor	Read	1h06min	Orthographic	1			Free	2018
West Point	Read	1h40min	Orthographic	128		LDC User Agreement for Non-Members	US$250 US$500	2008
Spoltech	Spontaneous and read	5h	Orthographic and phonetic	480	Uncontrolled	CSLU Agreement	US$150	2006
Sid	Spontaneous and read	4h57min	Orthographic	72	Uncontrolled			

Table 2 – Speech Corpora for ASR in PT-BR

In Table 2, blank cells indicate non-available data. This table underscores the scarcity of speech corpora with readily available phonetic transcriptions in PT-BR, as only two — Spoltech and MF — of the selected corpora provide this data. MF (Batista; Dias; Neto, 2022) provides not only orthographic and phonetic transcriptions but also time stamps for forced phonetic alignment, however it is only 15 min long and features one female and one male speaker. The West Point corpus offers a pronunciation dictionary for generating automatic phonetic transcriptions from the orthographic transcriptions. The lack of further resources for APT for PT-BR corroborates the need for a freely accessible dataset with phonetic transcriptions aimed at APT for this language.

Junior et al. (2023), authors of the CORAA ASR corpus, emphasize that PT-BR ASR models are commonly trained with an ensemble of Common Voice, Sid, VoxForge, and LapsBM — datasets featuring non-conversational speech. They also stress the importance of incorporating speech of various genres, from interviews to informal dialogues and conversations, in training robust ASR systems. This is because ASR systems trained on read style and clean speech (or even on prepared speech) often face performance drops when confronted with informal conversations in dynamic and noisy settings. Hence, it is imperative to leverage speech corpora containing spontaneous speech in uncontrolled environments to enhance ASR system applicability across diverse contexts. It is also crucial to include as many varieties of speakers as possible, aiming to increase accessibility of the ASR system. For this reason, we prioritized the CORAA ASR corpus for our work.

3.1.1 CORAA ASR

CORAA ASR encompasses five validated corpora (ALIP, C-ORAL-BRASIL I, NURC Recife, SP2010, and TEDx Portuguese talks) adapted for the task of ASR (JUNIOR et al., 2023). The corpus is divided into three datasets — train (286.31 h), dev (5.76 h), and test (11.63 h), and covers four primary accents — São Paulo (state), São Paulo (capital), Minas Gerais, and Recife — and miscellaneous accents. Figure 4 shows the distribution of these accents in the corpus.

Most audios have a Recife accent, however the São Paulo and Minas Gerais accents comprise a substantial part of the dataset too. There are also a few audios in European Portuguese (PT-PT) but, since our study focuses on PT-BR, we opted not to include the PT-PT audio-transcriptions in our dataset.

Figure 5 displays the speech styles present in the training set. 74.37% of the audios contain spontaneous speech, while 14.08% contain prepared speech and 11.55% contain both spontaneous and read speech.

In addition, the CORAA corpus contains validated audio-transcription pairs. Through a simple web interface, annotators verified each audio-transcription pair, upvoting valid ones and downvoting those with problems. Six scenarios led annotators to reject a pair,

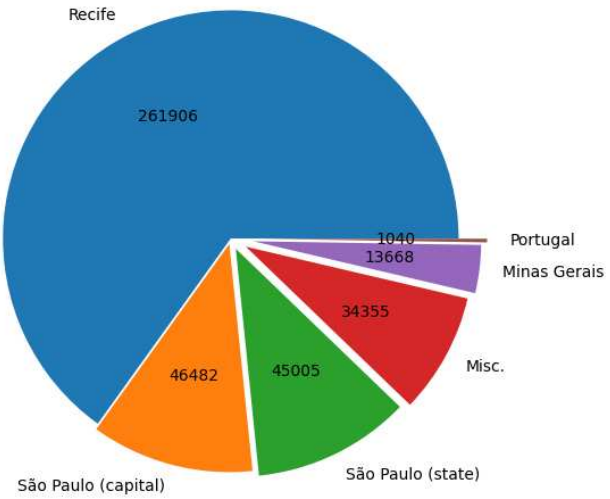


Figure 4 – Accent distribution.

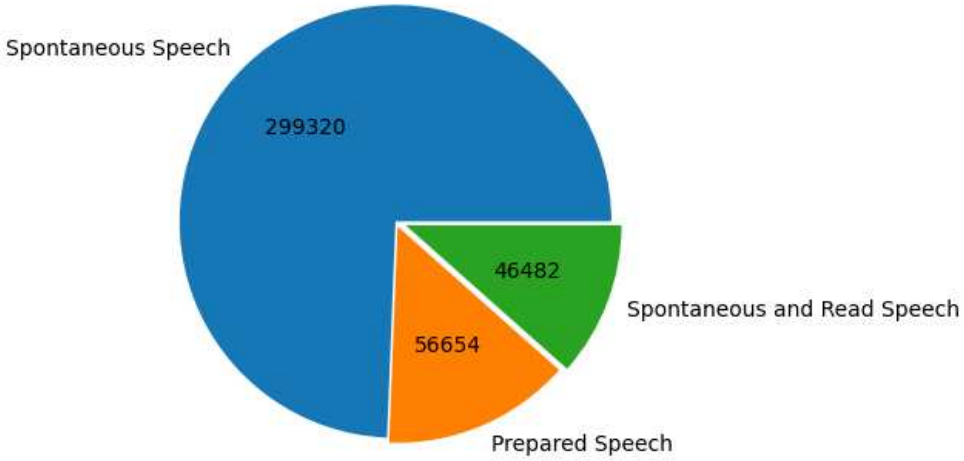


Figure 5 – Speech style distribution.

specifically:

1. Voice overlapping;
2. Low volume of the main speaker's voice, making the audio incomprehensible;
3. Word truncation;
4. Too many words in the transcript (compared to the audio);
5. Too few words in the transcript (compared to the audio);
6. Words swapped (transcript unaligned).

Moreover, five conditions determined whether a pair was considered valid, namely:

1. Valid without problems;
2. Valid with filled pause(s);
3. Valid with hesitation;
4. Valid with background noise/low voice but understandable;
5. Valid with little voice overlapping.

Table 3 shows that approximately 80.21% of audio-transcription pairs received at least one up vote, and 96.60% received no down votes— demonstrating high-quality standards. Additionally, no pair in the published corpus has more downvotes than upvotes. To select top-tier audio-transcription pairs, we kept those with at least one up vote (meaning they were reviewed at least once) and without any down votes, totaling 293,610 pairs or 76.81% of the data. This ensures greater accuracy and reliability in our dataset.

Number of Votes	Upvote	Downvote
0	19.98	96.53
1	64.64	3.40
2	7.60	0.07
3	7.78	0.00

Table 3 – Percentage of votes for 402456 audio-transcription pairs.

Table 4 shows the percentages of problems in the audios or lack thereof (no identified problem).

Number of Votes	Hesitation	Filled Pause	Noise Or Low	Second Voice	No Identified Problem
0	91.50	94.69	Voice	97.62	33.12
1	7.25	4.47	31.85	2.11	56.02
2	0.91	0.55	3.04	0.24	7.53
3	0.34	0.29	0.59	0.04	3.32

Table 4 – Percentage of votes for 382258 audio-transcription pairs.

Additionally, though 35.94% contain at least one identified problem, we decided to keep these audios in our dataset to reflect the original characteristics of the corpus. This will also allow for fine-tuning a more robust APT model, able to deal with minor interference in the audios.

Figure 6 shows the histogram of the duration of the filtered audios. The different corpora within CORAA ASR, with a large diversity of audio types, explain the variability in the duration.

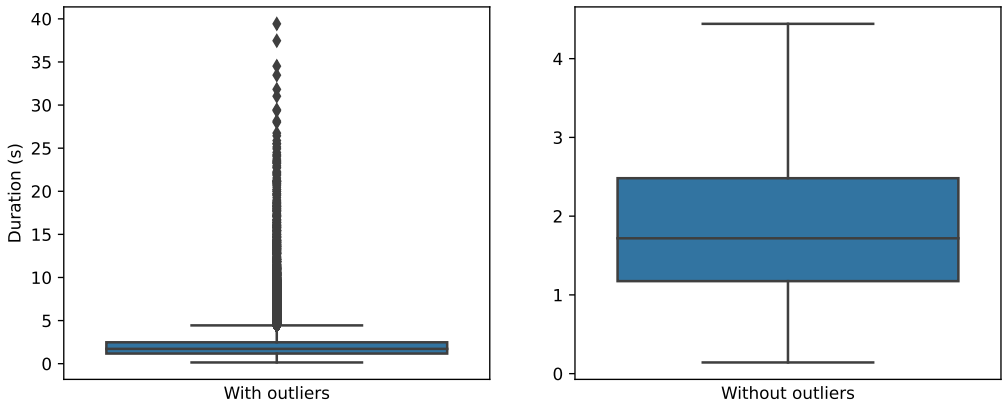


Figure 6 – Boxplots of the duration of the audios in the proposed dataset.

3.1.2 G2P

We tested several G2P converters for PT-BR on a dataset of manually verified words with phonetic transcriptions scraped from the "Portal da Língua Portuguesa", or "Portuguese Language Portal" in English. The portal is an organized repository of language resources aimed at both the general public and the scientific community (CORREIA; ASHBY; JANSSEN, 2010). The Portal has a phonetic dictionary of nearly 53,400 words in 10 Portuguese accents, including Luanda, Rio de Janeiro, São Paulo, Maputo, Lisbon, and Díli). We adopted the standard São Paulo phonetic transcriptions as reference transcriptions because we found that this particular PT-BR accent was more generalizable to other regions of Brazil.

word	category	transcription
a	artigo	a
á-bê-cê	nome masculino	ˌa.bˌe.sˈe
a-pro · pó · si · to	nome masculino	a.pɾo.pˈɔ.zi.tʊ
à-von · ta · de	nome masculino	ˌa.võ.tˈa.dʒi
a · a	nome feminino	a.ˈa

Table 5 – Head of the Portuguese Language Portal's dataset

As shown in Table 5, the original data included primary and secondary stress marks and special characters. These markers were removed from the transcriptions. We also used an iterative process of transcribing the dataset with the G2P converters, analyzing the words with the highest PERs, and manually correcting any faulty transcriptions identified in the dataset. In this process, some inaccuracies that contributed to a higher PER were corrected, such as in the examples shown in Table 6.

The choice of the G2P converters was made by extensively investigating the literature and online repositories such as GitHub and GitLab.

Error	Frequency
Duplicated transcriptions	686
Transcriptions that consist only of /ou/	26
Wrong or missing suffix	3
Transcriptions that correspond to other words	2
Transcription with an invalid phoneme	1

Table 6 – Errors identified in the Portuguse Language Portal's database

FalaBrasil G2P was presented by Neto et al. (2011), Silva et al. (2006) as a rule-based G2P converter with stress determination for PT-BR, which has self-contained rules that do not rely on other algorithms such as syllabic division or plural identification.

eSpeak (DUDDINGTON, 1995) and eSpeak-NG (DUNN, 2015) are both multilingual and open source software TTS synthesizers for Linux and Windows. They use a phonemization rule engine for each of the supported languages, which include PT-BR and PT-PT. The eSpeak-NG project emerged with the intention of cleaning up the existing eSpeak codebase, adding new features, and improving the supported languages (DUNN, 2015).

Novak, Minematsu e Hirose (2016) introduced Phonetisaurus, an open-source G2P conversion toolkit that leverages the OpenFst library. Phonetisaurus is a data-driven G2P trained with a variation of the joint multigram approach with the Weighted Finite-State Transducer paradigm. The Phonetisaurus repository contains a PT-BR model trained on a pronunciation dictionary.

To conclude, Epitran is a multilingual, multiple back-end system for G2P transduction which is distributed with support for 61 languages. Epitran (MORTENSEN; DALMIA; LITTELL, 2018) uses a mapping-and-repairs approach to G2P. It is expected that there is a mapping between graphemes and phonemes that can do most of the work of converting orthographic representations to phonological representations (HULDEN; GORMAN; LEE, 2014). For this reason, Epitran works best with phonetically consistent languages.

The G2P converters were evaluated as to the discordance rate between the generated automatic transcriptions and the reference transcriptions. This rate, calculated as the PER, embed variations in phonetic inventories, accents, representations of certain phoneme groups, and any potential errors in the generated G2P transcriptions, although these are difficult to precisely identify. The original vocabularies of each G2P tool, as well as the vocabulary of the reference phoneme charts and of the phonetic dictionary, are included in Appendix B. To better evaluate the discordance rates, we selected a random 10% sample of the dataset, i.e. 5337 words, that is enough to represent the population.

In addition, in order to better compare the G2P quality, we transformed the vocabularies of the dictionary and the G2P converters according to the reference phoneme

charts discussed in Section *2.1.3.2*. Furthermore, even though the charts do not include the approximants /j/ and /w/, we kept these phonemes wherever they appeared because they are present in the examples included in Ivo (2019a) and Ivo (2019b). The transformed vocabularies, as well as the rules used in this process, are included in Appendix C.

In summary, we chose the G2P tool that yielded the lowest discordance rate to transcribe the CORAA ASR corpus. The only correction made in the transcriptions was the pronunciation of "aham", initially transcribed as /aãw̃/, then corrected to /ãχã/, after consultation with a student of Linguistics. The transcribed corpus was then used to fine-tune the wav2vec 2.0 model, as detailed in Section 3.2. The datasets, configuration files, and Python scripts are made available in **this GitHub repository**[1].

3.2 PT-BR APT model

To define the train, test and dev sets of the wav2vec 2.0 model, we selected audios with a duration between the first quartile (Q1) and the third quartile (Q3) represented in Figure 6, delimiting the interval [1.17, 2.48]. We also filtered out transcriptions that contained only one phoneme, such as "ah", "uh", and "eh", to avoid including audios with little voice activity, which would not add much information to the wav2vec 2.0 model.

Furthermore, we defined three subsets of the transcribed train set of the CORAA ASR, comprising 1 hour, 10 hours, and 60 hours of audio. The aim was to evaluate the impact of the train set size in the final result. The randomly selected test set and dev set comprised 1 hour of audio each.

There are a number of pre-trained multilingual wav2vec 2.0 models, such as XLS-R (RUDER; SØGAARD; VULIĆ, 2019) and XLSR-53 (BABU et al., 2022), which can undergo fine-tuning in order to be used in APT tasks. We used XLSR-53 and we followed the tutorials by Platen (2021) and Kitahara (2021) to prepare the dataset, the tokenizer, and the feature extractor, and to perform the fine-tuning. The model vocab was encoded using a custom encoding in order to keep the tokens with a length of 1, which facilitated the evaluation of the model. The best fine-tuned model is shared in the **Hugging Face repository**[2]. The graphs showing the fine-tuning progress are included in Appendix D. We also share the training configurations in Appendix E but they are also included in the GitHub repository.

[1] <https://github.com/caiocrocha/Brazilian_Quick_APT>. Accessed: July 2, 2024.

[2] <https://huggingface.co/caiocrocha/wav2vec2-large-xlsr-53-phoneme-portuguese>. Accessed: July 2, 2024.

4 Results

In this chapter, we present the results and discussion of the selection process of the G2P converter and of the fine-tuning of the PT-BR APT model.

The experiments were divided in two parts. In the first part, we analyzed the performance of the selected G2P tools on the Portuguese Language Portal's dataset. In Section 4.1, we present the discordance rates in relation to the reference transcriptions and we investigate transcription errors. In the second part, we fine-tuned the wav2vec 2.0 model on the transcribed CORAA datasets. In Section 4.2, we analyze the performance of the fine-tuned wav2vec 2.0 models and we examine the confusion between phonemes.

4.1 G2P

The discordance rates for each G2P converter before and after the transformation according to the mapping detailed in Appendix C are presented in Table 7.

G2P	Discordance rate before	Discordance rate after
eSpeak	0.3076	0.2550
eSpeak-NG	0.3075	0.2548
FalaBrasil	**0.2980**	**0.2257**
Epitran	0.5237	0.4904
Phonetisaurus	0.4332	0.4020

Table 7 – Discordance rates before and after the transformation of the G2P transcriptions.

FalaBrasil yielded the lowest discordance rate after the transformation of the vocabulary. Meanwhile, eSpeak-NG and eSpeak come in second and third and obtained very similar results because they differ only in the version of the G2P converter. Even though these three present modestly high discordance rates, we couldn't find any obvious errors in their outputs. On the other hand, Epitran and Phonetisaurus presented some evident errors and thus obtained much higher discordance rates. Epitran's transcriptions included the phoneme /k/ instead of /ʃ/ in words like "charuto" and "chalé". Phonetisaurus' transcriptions replaced /l/ with /w/ and /z/ with /s/ in words like "oleaginoso". In addition, both confused /r/ for /χ/ in words like "micro".

Another crucial factor that contributed to a higher discordance rate was the G2P accent. Though we couldn't find information on the accents adopted for each G2P tool, we observed that they utilized varied PT-BR accents. eSpeak and eSpeak-NG were likely constructed to generate transcriptions in the standard São Paulo accent. Epitran is likely representative of a Northeastern PT-BR accent, as it features the phoneme /ʃ/ in the plural of words ("S chiado") and the open-mid vowels /ɛ/ and /ɔ/ in words such as "nordestino". We couldn't determine the precise accent of FalaBrasil and Phonetisaurus, though they very likely represent a standard Southeastern accent. Because the Portuguese

Language Portal's database features the standard São Paulo accent, eSpeak, eSpeak-NG, and FalaBrasil took benefit from this. On the other hand, Epitran obtained higher discordance rates, though they were also due to the aforementioned transcription errors.

In addition, we calculated the discordance rate by word for the G2P tools. Figure 7 shows these discordance rates for FalaBrasil and eSpeak-NG.

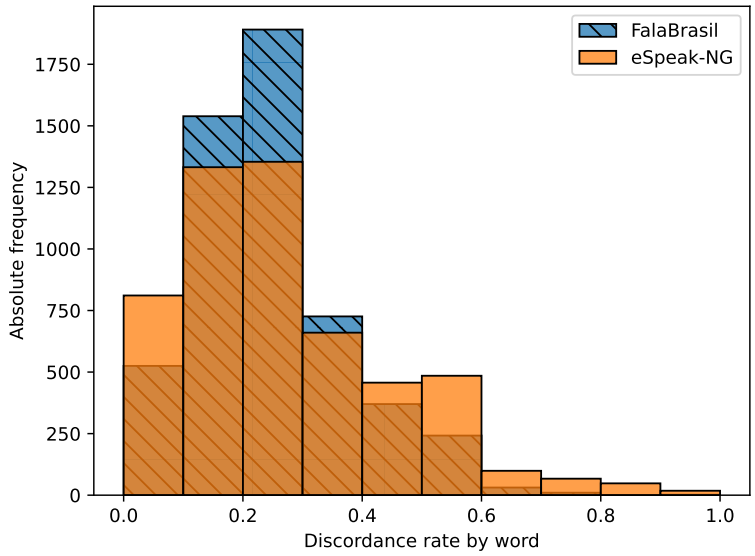


Figure 7 – Histogram of the discordance rates by word for FalaBrasil and for eSpeak-NG.

The histogram shows that FalaBrasil had a peak of discordance rate in the range [0.1, 0.3] but it presented very few words with a discordance rate higher than 0.6. On the other hand, eSpeak-NG had a more even distribution of discordance rates, stretching up to 1.0, meaning that it had more words with a very low discordance rate but also words with a very high discordance rate.

4.2 PT-BR APT model

We fine-tuned three models with 1 hour, 10 hours, and 60 hours of audio each, which we named XLSR-APT-1h, XLSR-APT-10h, and XLSR-APT-60h respectively. The PERs for each model are presented in Table 8.

Model \ Dataset	Dev	Test
1h	0.8301	0.7963
10h	0.2197	0.1587
60h	0.2190	0.1600

Table 8 – PER of the fine-tuned models on each set.

XLSR-APT-10h achieved the lowest PER on the test set, indicating that 10 hours of audio were sufficient for satisfactory performance. Adding 50 more hours to the train set did not improve the model, as shown by XLSR-APT-60h's PER. In contrast, XLSR-APT-1h had a PER 400% higher, highlighting its significantly poorer performance.

Table 9 displays the accuracy of each model, showing how well the predicted phonemes matched the expected ones.

Model \ Dataset	Dev	Test
1h	0.58	0.58
10h	0.89	0.92
60h	0.89	0.92

Table 9 – Accuracy of the fine-tuned models on each set.

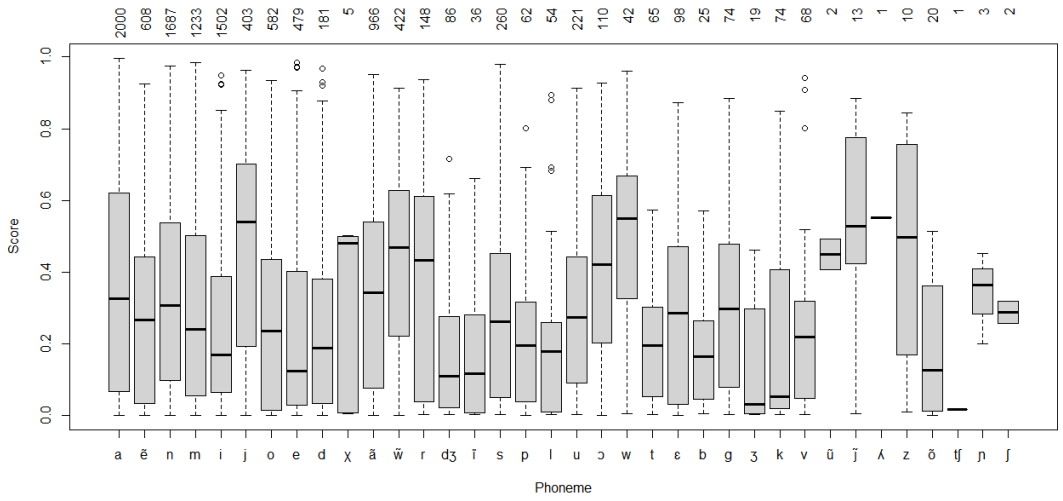
XLSR-APT-1h had a low accuracy, confirming its poor performance. In contrast, XLSR-APT-10h and XLSR-APT-60h achieved a reasonable accuracy, indicating acceptable performances. Sections **4.2.1** and **4.2.2** provide a more in-depth analysis of the performance of each model.

4.2.1 Confidence analysis

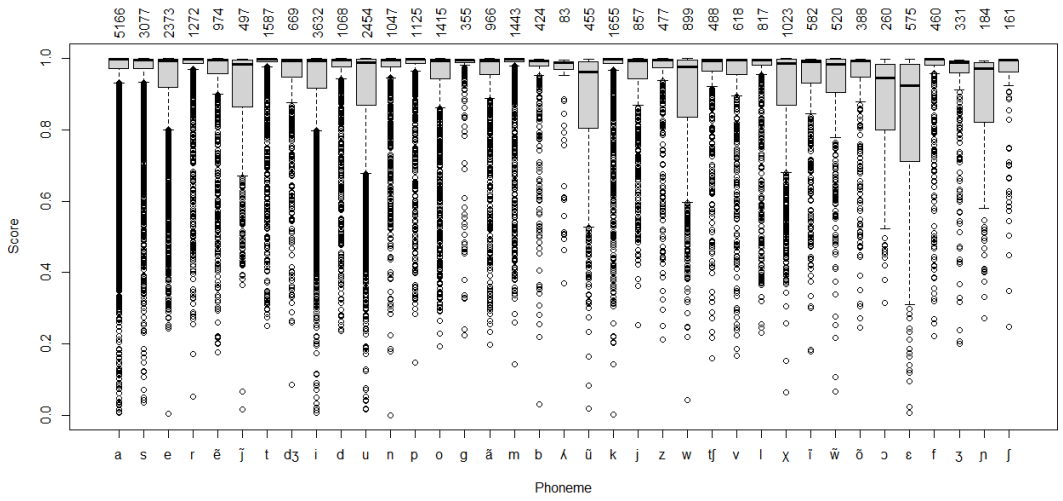
To get a better sense of the models' predictions, we plotted box plots of the confidence scores by predicted phoneme, shown in Figure 8.

Figure 8a shows that XLSR-APT-1h presented little confidence in its predictions. In addition, the support of the predicted phonemes is lower than in the reference transcriptions, which indicates that the model performed poor transcriptions. This is a very evident hint that the model did not properly learn the phonemes in the train set.

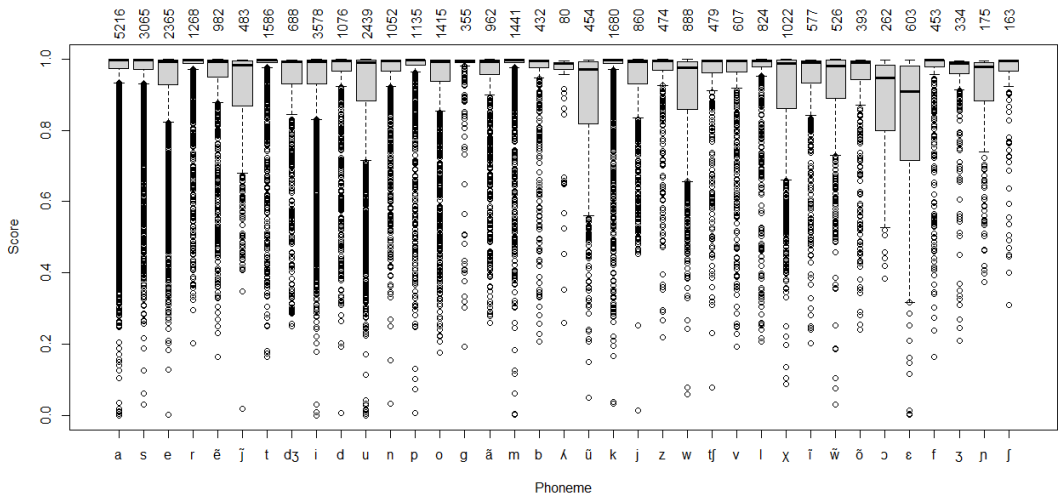
Figures 8b and 8c show that, for the XLSR-APT-10h and XLSR-APT-60h models, while some phonemes exhibited higher variance in confidence scores, the median confidence level, denoted by the black horizontal line in each bar, remained consistently high across all phonemes. This suggests overall confidence in the models' predictions. However, the considerable variance in certain phonemes indicates challenges in recognizing specific utterances. Potential reasons for this include errors in the G2P output, unclear or muffled speech, and external noise.



(a) XLSR-APT-1h



(b) XLSR-APT-10h

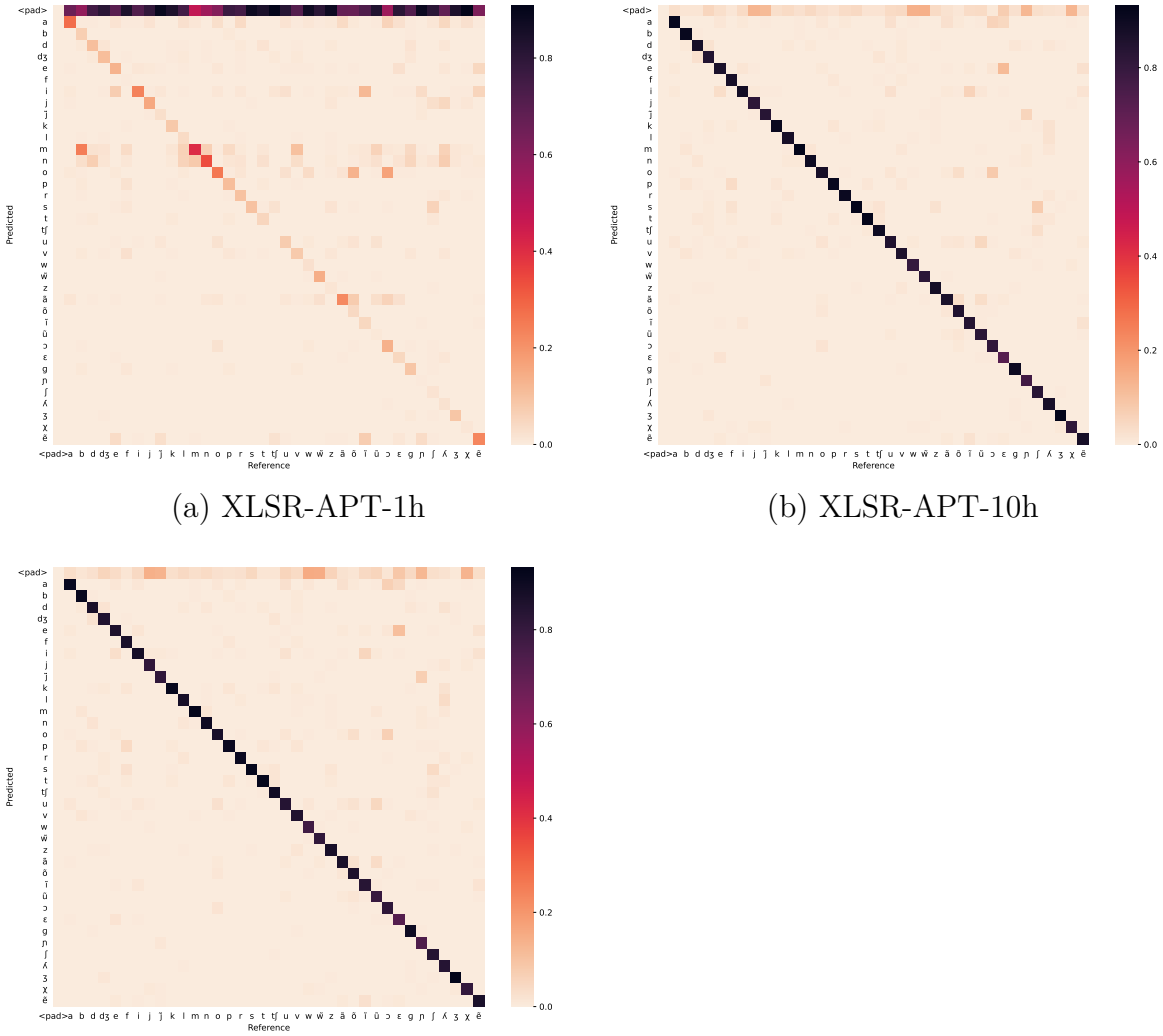


(c) XLSR-APT-60h

Figure 8 – Boxplot of the confidence scores by predicted phoneme for the test set.

4.2.2 Confusion analysis

The confusion matrices presented in Figure 9 provide a detailed view of the models' accuracy, showing the correct predictions in the diagonal and the wrong predictions in the off-diagonal cells. "<pad>" is the padding token used as CTC-blank label, representing predictions where the models could not identify any specific phoneme.



(a) XLSR-APT-1h

(b) XLSR-APT-10h

(c) XLSR-APT-60h

Figure 9 – Confusion matrices for the test set.

Figure 9a shows that XLSR-APT-1h confused all phonemes with "<pad>", exposing yet again that the model failed to predict the phonemes in the test set. On the other hand, Figures 9b and 9c depict a much improved performance, with few confusions between phonemes. However, they are very similar, revealing that XLSR-APT-60h did not bring any noticeable improvements over XLSR-APT-10h. When analyzing the phonemes with highest confusion, we can observe confusions between similar vowels, such as /ɛ/ and /e/, /ɔ/ and /o/, /ũ/ and /u/, /ĩ/ and /i/. Additionally, close consonants are also confounded, such as /ʃ/ and /s/, /ɲ/ and /j̃/. At last, some phonemes, such as /j/, /j̃/, /w/, /w̃/, /ɲ/,

and /χ/, were mistaken for "<pad>", indicating that the models had trouble recognizing them.

5 Conclusion

This research aimed to explore a methodology to create a dataset of phonetic transcriptions for PT-BR from existing ASR corpora. Based on a thorough literature review of PT-BR speech corpora, we analyzed their extensiveness, the diversity of accents, their fitness to APT model training, their price, and their license. We discovered a lack of APT resources tailored for PT-BR, as highlighted by Aguiar e Costa-Abreu (2023), Dijkstra (2021). The main speech corpora found to have human-validated phonetic transcriptions were MF (Batista; Dias; Neto, 2022) and Spoltech (SCHRAMM et al., 2006). However, these speech corpora are either small, as is the case of MF, or paid, such as Spoltech. On the other hand, several speech corpora are large enough and present text transcriptions that can be used for the training of ASR models.

We chose CORAA ASR (JUNIOR et al., 2023) to develop a dataset of phonetic transcriptions derived with FalaBrasil's G2P converter (NETO et al., 2011). The choice of the G2P tool was made by analyzing the G2P output compared to the Portuguese Language Portal's dictionary (CORREIA; ASHBY; JANSSEN, 2010) of phonetic transcriptions. We also adopted a reference phoneme vocabulary (IVO, 2019a; IVO, 2019b) and standardized the phonetic transcriptions (Appendix C). FalaBrasil's G2P presented the lowest discordance rate among the selected G2Ps. Moreover, we concluded that standardizing the phonetic transcriptions reduced the phoneme discordance rate by up to 24%.

The transcribed dataset of phonetic transcriptions derived from CORAA ASR was used to fine-tune three XLSR-53 wav2vec 2.0 models (RUDER; SØGAARD; VULIĆ, 2019) for APT in PT-BR. We found that the model fine-tuned on 10 hours of audio achieved satisfactory performance, obtaining a 15.87% PER on the test set. By analyzing the confidence scores per class, we observed that it presented high mean confidence scores. However, there were some exceptions, such as /j/, /j̃/, /w/, /w̃/, /ɲ/, and /χ/, which had a higher confidence score variance. Similarly, by analyzing the confusion matrix, we noted that /ɛ/, /e/, /ɔ/, /o/, /ũ/, /u/, /ĩ/, /i/, /ʃ/, /s/, /ɲ/ and /j̃/ presented confusion rates up to 20%, indicating that the prediction accuracy could still be improved.

One reason for such confusions might be the difference between the G2P transcriptions used for training and the actual uttered phonemes. The training transcriptions derived with FalaBrasil's G2P account for one accent but the audios include several accents, such as the Recife, Minas Gerais, standard and non-standard São Paulo accents. In consequence, the model has certainly encountered different pronunciations of the same words. Since wav2vec 2.0 jointly learns contextualized speech representations and an inventory of discretized speech units (BAEVSKI et al., 2020) (phones), these different phonemic representations for the same contexts (words) coupled with non-matching G2P

transcriptions may have introduced confusion during the fine-tuning process. In fact, /ʃ/ and /s/, /ɛ/ and /e/, /ɔ/ and /o/ are allophones in PT-BR that vary according to the accent. The fact that, according to Figure 4, 65% of the audios in the corpus feature the Recife accent, characterized by the pronunciation of /ʃ/, /ɛ/, and /ɔ/ in certain contexts, contrarily to the accent featured in FalaBrasil's G2P accent, which presents rather /s/, /e/, /o/ for the same contexts, also supports this hypothesis.

One solution to this problem could involve using a different G2P tool for each accent featured in the corpus, increasing the likelihood that the phonetic transcriptions match the actual utterances and reducing confusion. Additionally, optimizing the fine-tuning hyperparameters could further enhance the outcome.

Additionally, the study by Bhatt et al. (2023) underscored that researchers used triphone-based context-dependent phonemes to reduce the contextual effect, improving phoneme recognition. The authors also stated that identifying voicing is essential for consonant recognition, because there is often an important distinction between voiced and unvoiced stops, as well as voiced and unvoiced fricatives. Similarly, identifying nasalization is crucial for accurate vowel recognition. Therefore, increasing the representation of both unvoiced and voiced stops, as well as voiced and unvoiced fricatives and nasal vowels in the training set, could lead to improved phoneme recognition.

To conclude, we highlight the potential use of this work for phoneme recognition applications in PT-BR, such as speech therapy, educational assessment, language learning, linguistic research, and the development of ASR technology.

REFERENCES

AGUIAR, T.; COSTA-ABREU, M. D. Automatic collection of transcribed speech for low resources languages. In: *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*. [S.l.: s.n.], 2023. p. 1–6.

ALHARBI, S. et al. Automatic speech recognition: Systematic literature review. *IEEE Access*, v. 9, p. 131858–131876, 2021.

BABU, T. A. et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In: . [S.l.: s.n.], 2022. p. 2278–2282.

BAEVSKI, A.; CONNEAU, A.; AULI, M. *Wav2vec 2.0: Learning the structure of speech from raw audio*. 2020. Accessed November 21 2023. Disponível em: <https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/>.

BAEVSKI, A. et al. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020. (NIPS'20). ISBN 9781713829546.

Batista, C.; Dias, A. L.; Neto, N. Free resources for forced phonetic alignment in Brazilian Portuguese based on Kaldi toolkit. *EURASIP Journal on Applied Signal Processing*, v. 2022, n. 1, p. 11, dez. 2022.

BHATT, S. et al. A comprehensive examination of phoneme recognition in automatic speech recognition systems. *Traitement du Signal*, v. 40, n. 5, p. 1997–2008, 2023.

BOSCH, A. V. D.; DAELEMANS, W. Data-oriented methods for grapheme-to-phoneme conversion. In: *Sixth Conference of the European Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 1993.

CORREIA, M.; ASHBY, S.; JANSSEN, M. *Portal da Língua Portuguesa*. Avenida Elias Garcia, 147, 5.º Dt.º, 1050-099 Lisbon (Portugal): Instituto de Linguística Teórica e Computacional (ILTEC), 2010. <http://www.portaldalinguaportuguesa.org>.

CUCCHIARINI, C.; STRIK, H. Automatic phonetic transcription: An overview. In: CITESEER. *Proceedings of ICPHS*. [S.l.], 2003. p. 347–350.

DENG, L. Deep learning: from speech recognition to language and multimodal processing. *APSIPA Transactions on Signal and Information Processing*, v. 5, 01 2016.

DIJKSTRA, B. A. *Phoneme Recognition using Convolutional Neural Networks for Automatic Phonetic Transcription*. Dissertação (Mestrado) — Federal University of Technology – Parana, Ponta Grossa, 2021.

DUDDINGTON, J. *eSpeak: Speech Synthesizer*. 1995. <http://espeak.sourceforge.net/>. Accessed: April 19, 2024.

DUNN, R. *eSpeak NG: Next Generation Speech Synthesizer*. 2015. <https://github.com/espeak-ng/espeak-ng>. Accessed: April 19, 2024.

ENGELHART, E.; ELYASI, M.; BHARAJ, G. *Grapheme-to-Phoneme Transformer Model for Transfer Learning Dialects*. 2021.

GANDHI, S. et al. *Hugging Face Audio Course - Chapter 5: Evaluation*. 2023. <https://huggingface.co/learn/audio-course/chapter5/evaluation>. Accessed on 11 December 2023.

GRIS, L. R. S. et al. Brazilian portuguese speech recognition using wav2vec 2.0. In: PINHEIRO, V. et al. (Ed.). *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2022. p. 333–343. ISBN 978-3-030-98305-5.

GROSMAN, J. *Fine-tuned XLSR-53 large model for speech recognition in Portuguese*. 2021. <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-portuguese>.

GROSMAN, J. *Fine-tuned XLS-R 1B model for speech recognition in Portuguese*. 2022. <https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-portuguese>.

HULDEN, M.; GORMAN, K.; LEE, J. *Epitran*. 2014. <https://github.com/dmort27/epitran>. Accessed: April 19, 2024.

IVO, A. *Aula 02 Fonética Articulatória (Consoantes do português brasileiro)*. 2019. <https://grad.letras.ufmg.br/arquivos/monitoria/Aula%2002%20apoio.pdf>. Apoio Pedagógico – Estudos Linguísticos: Fonética e Fonologia.

IVO, A. *Aula 03: Fonética Articulatória (Vogais do português brasileiro e treino de transcrições fonéticas)*. 2019. <https://grad.letras.ufmg.br/arquivos/monitoria/Aula%2003%20apoio.pdf>. Apoio Pedagógico – Estudos Linguísticos: Fonética e Fonologia.

JOUVET, D.; FOHR, D.; ILLINA, I. Evaluating grapheme-to-phoneme converters in automatic speech recognition context. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2012. p. 4821–4824.

JUNIOR, A. C. et al. CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *Language Resources and Evaluation*, 2023.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. [S.l.: s.n.], 2023. Copyright © 2023. All rights reserved. Draft of January 7, 2023.

KITAHARA, K. *Fine-tuning XLSR-Wav2Vec2 for Phoneme Recognition with Transformers*. 2021. <https://github.com/kosuke-kitahara/xlsr-wav2vec2-phoneme-recognition/blob/main/Fine_tuning_XLSR_Wav2Vec2_for_Phoneme_Recognition.ipynb>. Last Updated Mar 29.

KLUMPP, P. et al. Common phone: A multilingual dataset for robust acoustic modelling. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022. p. 763–768. Disponível em: <https://aclanthology.org/2022.lrec-1.81>.

KUNAPULI, S. S.; BHALLAMUDI, P. C. Chapter 22 - a review of deep learning models for medical diagnosis. In: KUMAR, P.; KUMAR, Y.; TAWHID, M. A. (Ed.). *Machine Learning, Big Data, and IoT for Medical Informatics*. Academic Press, 2021, (Intelligent Data-Centric Systems). p. 389–404. ISBN 978-0-12-821777-1. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128217771000070>.

LI, J. et al. *Robust automatic speech recognition: A bridge to practical applications*. [S.l.: s.n.], 2015. 1-286 p.

LI, X. et al. Universal phone recognition with a multilingual allophone system. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8249–8253, 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:211532368>.

LI, X. et al. Hierarchical phone recognition with compositional phonetics. In: *Interspeech*. [S.l.: s.n.], 2021. p. 2461–2465.

LOPES, C.; PERDIGAO, F. Phoneme recognition on the timit database. In: IPSIC, I. (Ed.). *Speech Technologies*. Rijeka: IntechOpen, 2011. cap. 14. Disponível em: <https://doi.org/10.5772/17600>.

MALAKAR, M.; KESKAR, R. B. Progress of machine learning based automatic phoneme recognition and its prospect. *Speech Communication*, v. 135, p. 37–53, 2021. ISSN 0167-6393. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167639321001084>.

MEHRISH, A. et al. A review of deep learning techniques for speech processing. *Information Fusion*, v. 99, p. 101869, 2023. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253523001859>.

MIHAJLIK, P.; RÉVÉSZ, T.; TATAI, P. Phonetic transcription in automatic speech recognition. *Acta Linguistica Hungarica*, Akadémiai Kiadó, v. 49, n. 3-4, p. 407–425, 2002.

MITTERER, H.; CUTLER, A. Speech perception. In: BROWN, K. (Ed.). *Encyclopedia of Language and Linguistics (Second Edition)*. Second edition. Oxford: Elsevier, 2006. p. 770–782. ISBN 978-0-08-044854-1. Disponível em: <https://www.sciencedirect.com/science/article/pii/B0080448542000298>.

MOHAMED, A. rahman et al. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, v. 16, p. 1179–1210, 2022. Disponível em: <https://api.semanticscholar.org/CorpusID:248987289>.

MORTENSEN, D. R.; DALMIA, S.; LITTELL, P. Epitran: Precision G2P for many languages. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1429>.

NASSIF, A. B. et al. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, v. 7, p. 19143–19165, 2019.

NEDJAH, N.; BONILLA, A. D.; de Macedo Mourelle, L. Automatic speech recognition of portuguese phonemes using neural networks ensemble. *Expert Systems with Applications*, v. 229, p. 120378, 2023. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417423008801>.

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>.

NOVAK, J. R.; MINEMATSU, N.; HIROSE, K. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework. *Natural Language Engineering*, v. 22, n. 6, p. 907–938, 2016.

O'SHAUGHNESSY, D. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, v. 41, n. 10, p. 2965–2979, 2008. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320308001799>.

PLATEN, P. von. Fine-tuning wav2vec 2.0 for english asr with transformers. *Hugging Face Blog*, March 2021. Disponível em: <https://huggingface.co/blog/fine-tune-wav2vec2-english>.

QUINTANILHA, I. M. *End-to-End Speech Recognition Applied to Brazilian Portuguese Using Deep Learning*. Dissertação (Mestrado) — Alberto Luiz Coimbra Institute for Graduate Studies and Engineering Research — Federal University of Rio de Janeiro, 03 2017.

RABINER, L. R.; SCHAFER, R. W. *Introduction to Digital Speech Processing*. Hanover, MA, USA: Now Publishers Inc., 2007. ISBN 1601980701.

RASO, T.; MELLO, H. *C-ORAL-BRASIL: corpus de referência do português brasileiro falado informal. I.* [S.l.]: Editora UFMG, 2012.

RAVANELLI, M. et al. Multi-task self-supervised learning for robust speech recognition. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2020. p. 6989–6993.

RICHEY, C. *Speech Corpora*. 2020. Accessed 19 November 2023. Disponível em: <https://web.stanford.edu/dept/linguistics/corpora/material/X_Speech_Corpora.pdf>.

RIZWAN, M.; ANDERSON, D. V. Comparison of distance metrics for phoneme classification based on deep neural network features and weighted k-nn classifier. In: *In Workshop on Machine Learning in Speech and Language Processing*. [S.l.: s.n.], 2016.

RUDER, S.; SØGAARD, A.; VULIĆ, I. Unsupervised cross-lingual representation learning. In: NAKOV, P.; PALMER, A. (Ed.). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Florence, Italy: Association for Computational Linguistics, 2019. p. 31–38. Disponível em: <https://aclanthology.org/P19-4007>.

SANTOS, F. W. dos; BARONE, D. A. C.; ADAMI, A. G. Validação de corpus para reconhecimento de fala contínua em português brasileiro. In: *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*. New York, NY, USA: Association for Computing Machinery, 2008. (WebMedia '08), p. 316–320. ISBN 9788576691990. Disponível em: <https://doi.org/10.1145/1809980.1810058>.

SAR, V.; TAN, T.-P. Applying linguistic g2p knowledge on a statistical grapheme-to-phoneme conversion in khmer. *Procedia Computer Science*, v. 161, p. 415–423, 2019. ISSN 1877-0509. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050919318502>.

SCHRAMM, M. C. et al. *CSLU: Spoltech Brazilian Portuguese Version 1.0*. Philadelphia: [s.n.], 2006. Web Download. LDC Catalog No.: LDC2006S16, ISBN: 1-58563-383-6, ISLRN: 386-396-917-783-5, Release Date: April 17, 2006, Member Year(s): 2006, DCMI Type(s): Sound, Text, Sample Type: 1-channel pcm, Sample Rate: 44100, Data Source(s): microphone speech, Application(s): language identification, language modeling, language teaching, machine learning, machine translation, Language(s): Portuguese, Language ID(s): por, License(s): CSLU Agreement, Online Documentation: LDC2006S16 Documents, Licensing Instructions: Subscription & Standard Members, and Non-Members.

SILVA, D. C. et al. A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing. In: *2006 International Telecommunications Symposium*. [S.l.: s.n.], 2006. p. 550–554.

TAGUCHI, C. et al. Universal automatic phonetic transcription into the international phonetic alphabet. In: University of Notre Dame, United States; Nara Institute of Science and Technology, Japan; Google, United States. *INTERSPEECH*. Dublin, Ireland, 2023.

TIWARI, A. Chapter 2 - supervised learning: From theory to applications. In: PANDEY, R. et al. (Ed.). *Artificial Intelligence and Machine Learning for EDGE Computing*. Academic Press, 2022. p. 23–32. ISBN 978-0-12-824054-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128240540000265>.

Van Bael, C. et al. Automatic phonetic transcription of large speech corpora. *Computer Speech and Language*, v. 21, n. 4, p. 652–668, 2007. ISSN 0885-2308. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0885230807000228>.

WELLS, J. C. Computer-coding the ipa: a proposed extension of sampa. *Revised draft*, Citeseer, v. 4, n. 28, p. 1995, 1995.

XU, Q.; BAEVSKI, A.; AULI, M. Simple and effective zero-shot cross-lingual phoneme recognition. In: META AI. *Interspeech*. Incheon, Korea, 2022.

YEH, C.-F. et al. Transformer-transducer: End-to-end speech recognition with self-attention. *arXiv preprint arXiv:1910.12977*, 2019.

YI, C. et al. *Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages*. 2021.

APPENDIX A – X-SAMPA to IPA

l	l
u	u
k	k
a	a
s	s
S	ʃ
r	r
t	t
m	m
i	i
o	o
f	f
w	w
Z	ʒ
e∼	ẽ
j∼	j̃
R	ʁ
e	e
j	j
i∼	ĩ
n	n
z	z
v	v
a∼	ã
w∼	w̃
E	ɛ
b	b
X	χ
d	d
dZ	dʒ
p	p
O	ɔ
g	ɡ
o∼	õ
tS	tʃ
u∼	ũ
J	ɲ
L	ʎ

Table 10 – X-SAMPA codes (left) and IPA symbols (right).

APPENDIX B – Phoneme vocabularies

The vocabularies of the reference phoneme charts, the Portuguse Language Portal, and the G2P converters are presented below:

Reference phoneme charts: ['p', 'b', 't', 'd', 'k', 'g', 'tʃ', 'dʒ', 'f', 'v', 's', 'z', 'ʃ', 'ʒ', 'χ', 'ɣ', 'h', 'ɦ', 'm', 'n', 'ɲ', 'ɾ', 'r', 'ɹ', 'l', 'ʎ', 'i', 'ĩ', 'ɪ', 'e', 'ẽ', 'ɛ', 'ə', 'ã', 'a', 'ʊ', 'ũ', 'u', 'õ', 'o', 'ɔ']

Portuguese Language Portal: ['a' 'b' 'e' 's' 'p' 'ɾ' 'o' 'ɔ' 'z' 'i' 't' 'ʊ' 'v' 'õ' 'dʒ' 'w' 'x' 'tʃ' 'g' 'd' 'j' 'ə' 'k' 'ə̃' 'ʃ' 'ʒ' 'f' 'm' 'ẽ' 'n' 'u' 'l' 'r' 'ʎ' 'ɛ' 'ɛ̃' 'ɲ' 'ũ' 'ĩ' 'ɨ' 'ɐ̃' 'h' 'ŋ' 'ɬ' 'ɔ̃']

eSpeak: ['l' 'u' 'k' 'æ' 's' 'ʃ' 'a' 'ɾ' 't' 'ʊ' 'm' 'i' 'r' 'o' 'f' 'ʒ' 'e' 'ɪ' 'ŋ' 'x' 'w' 'n' 'z' 'v' 'ɐ̃' 'ɔ̃' 'b' 'ɛ' 'd' 'dʒ' 'y' 'j' 'p' 'ɔ' 'g' 'tʃ' 'ũ' 'ɲ']

FalaBrasil: ['l' 'u' 'k' 'a' 's' 'ʃ' 'r' 't' 'm' 'i' 'o' 'f' 'w' 'ʒ' 'ẽ' 'j̃' 'ʁ' 'e' 'j' 'ĩ' 'n' 'z' 'v' 'ã' 'w̃' 'ɛ' 'b' 'χ' 'd' 'dʒ' 'p' 'ɔ' 'g' 'õ' 'tʃ' 'ũ' 'ɲ' 'ʎ']

Epitran: ['l' 'u' 'k' 'ɐ' 'ʃ' 'ɾ' 't' 'o' 'm' 'i' 'ʁ' 'f' 'ʒ' 'ɛ̃' 'ɛ' 'j' 'n' 'z' 'v' 's' 'ɐ̃' 'w̃' 'b' 'e' 'd' 'kʷ' 'a' 'p' 'ɔ' 'g' 'j́' 'ĩ' 'w' 'gʷ' 'ũ' 'õ' 'j̃' 'dʒ' 'ẽ' 'ẃ' 'lʒ']

Phonetisaurus: ['ɬ' 'u' 'k' 'a' 's' 'ʃ' 'ʁ' 't' 'o' 'm' 'i' 'f' 'g' 'e' 'ʒ' 'n' 'v' 'ɐ̃' 'b' 'ɛ' 'd' 'p' 'ɔ' 'z' 'ɲ' 'ẽ' 'ã' 'ʎ' 'õ' 'ʁ' 'ɾ' 'dʒ' 'w']

The vocabularies after the transformation according to the reference phoneme charts are presented below:

Portuguese Language Portal: ['l' 'u' 'k' 'ə' 's' 'ʃ' 'a' 'ɾ' 't' 'ʊ' 'm' 'i' 'o' 'f' 'w' 'ʒ' 'e' 'ɲ' 'χ' 'j' 'n' 'z' 'v' 'ã' 'b' 'ẽ' 'ɛ' 'd' 'dʒ' 'ĩ' 'p' 'ɔ' 'g' 'õ' 'tʃ' 'ũ' 'r' 'ʎ' 'h']

eSpeak: ['l' 'u' 'k' 'a' 's' 'ʃ' 'ɾ' 't' 'ʊ' 'm' 'i' 'r' 'o' 'f' 'ʒ' 'e' 'ɪ' 'n' 'χ' 'w' 'z' 'v' 'ã' 'ɔ̃' 'b' 'ɛ' 'd' 'dʒ' 'j' 'p' 'ɔ' 'g' 'tʃ' 'ũ' 'ɲ']

FalaBrasil: ['l' 'u' 'k' 'a' 's' 'ʃ' 'r' 't' 'm' 'i' 'o' 'f' 'w' 'ʒ' 'ẽ' 'j̃' 'χ' 'e' 'j' 'ĩ' 'n' 'z' 'v' 'ã' 'w̃' 'ɛ' 'b' 'd' 'dʒ' 'p' 'ɔ' 'g' 'õ' 'tʃ' 'ũ' 'ɲ' 'ʎ']

Epitran: ['l' 'u' 'k' 'ɐ' 'ʃ' 'ɾ' 't' 'o' 'm' 'i' 'χ' 'f' 'ʒ' 'ẽ' 'ɛ' 'j' 'n' 'z' 'v' 's' 'ã' 'w̃' 'b' 'e' 'd' 'a' 'p' 'ɔ' 'g' 'j́' 'ĩ' 'w' 'ũ' 'õ' 'j̃' 'dʒ' 'lʒ']

Phonetisaurus: ['w' 'u' 'k' 'a' 's' 'ʃ' 'χ' 't' 'o' 'm' 'i' 'f' 'g' 'e' 'ʒ' 'n' 'v' 'ã' 'b' 'ɛ' 'd' 'p' 'ɔ' 'z' 'ɲ' 'ẽ' 'ʎ' 'õ' 'ɾ' 'dʒ']

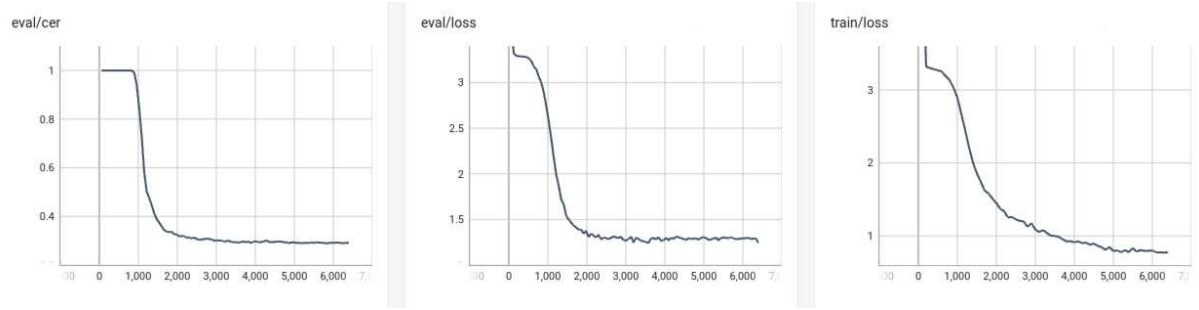
APPENDIX C – Transformation mapping

ŋ	→	n
ɨ	→	i
ə̃	→	ã
ɐ̃	→	ã
ɛ̃	→	ẽ
ɔ̃	→	õ
x	→	χ
g	→	ɡ
æ	→	a
ʁ	→	χ
ʀ	→	χ
ɡ^{w}	→	ɡ
k^{w}	→	k
ẃ	→	w
l^{j}	→	ʎ
y	→	i
ỹ	→	ɲ
ł	→	w

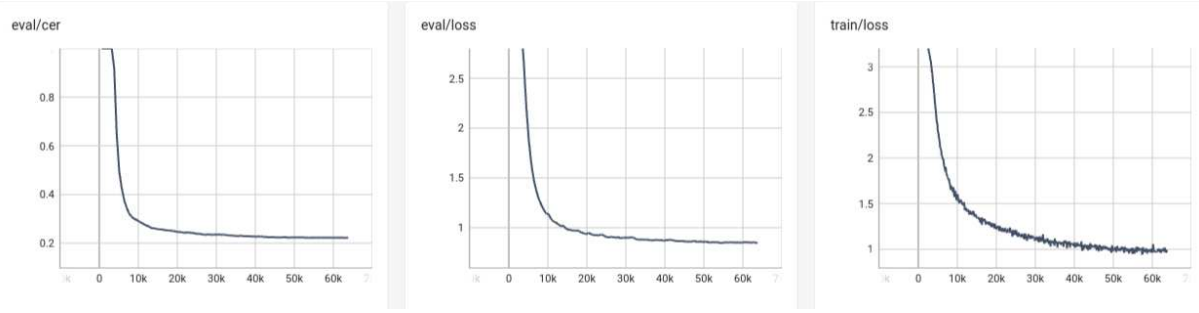
Table 11 – Mapping to transform the transcriptions according to the reference vocabulary.

APPENDIX D – Fine-tuning progress

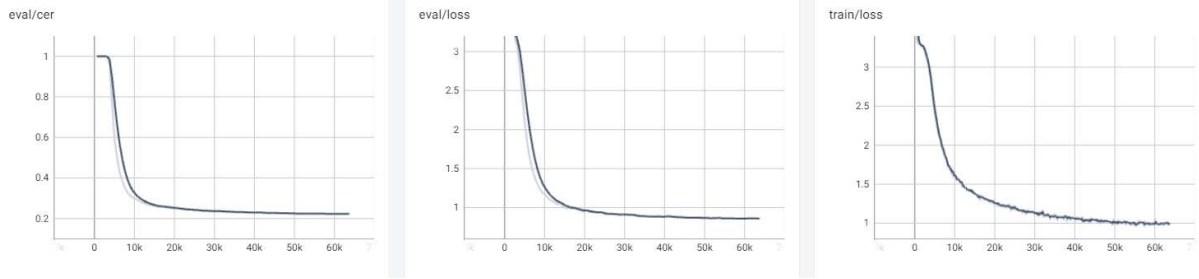
Figure 10 displays each models' fine-tuning progress.



(a) XLSR-APT-1h



(b) XLSR-APT-10h



(c) XLSR-APT-60h

Figure 10 – Fine-tuning progress.

APPENDIX E – Fine-tuning configuration

Listings 1, 2, and 3 include the *config.json* files used to fine-tune the wav2vec 2.0 models.


```
{
    "run_name": "Wav2Vec-fine-tuning-phonemes",
    "run_description": "Fine tuning phonemes",
    "seed": 42,

    "sampling_rate": 16000,

    "num_hidden_layers": 24,

    "vocab":{
        "vocab_path": "wav2vec2_phoneme_1h/vocab.json",
        "blank": "<pad>",
        "silence": "|",
        "unk": "<unk>"
    },

    "batch_size": 32,
    "mixed_precision": true,
    "early_stop_epochs": 50,

    "epochs": 100,
    "lr": 5e-5,
    "gradient_accumulation_steps": 1,

    "logging_steps": 100,
    "load_best_model_at_end": true,
    "save_total_limit": 2,
    "warmup_ratio": 0,
    "warmup_steps": 0,

    "num_loader_workers": 8,

    "freeze_feature_extractor": true,
    "attention_dropout": 0.1,
    "activation_dropout": 0.1,
```

```
"hidden_dropout": 0.1,
"feat_proj_dropout": 0.1,
"mask_time_prob": 0.1,
"layerdrop": 0.1,
"gradient_checkpointing": true,

"output_path": "wav2vec2_phoneme_1h/output",

"dataset_cache": "wav2vec2_phoneme_1h/datasets",

"datasets":{
    "train":
        [
          {
            "name": "csv",
            "path": "csv",
            "data_files":
            ↪  ["wav2vec2_phoneme_1h/input/metadata_train_final_g2p_ipa_sample_1h.csv"],
            "text_column": "transcript_encoded",
            "path_column": "file_path"
          }
        ]
    ,
    "devel":
        [

          {
            "name": "csv",
            "path": "csv",
            "data_files":
            ↪  ["wav2vec2_phoneme_1h/input/metadata_dev_final_g2p_ipa_sample_1h.csv"],
            "text_column": "transcript_encoded",
            "path_column": "file_path"
          }
        ]
        ,
    "test":
          {
            "name": "csv",
```

```
            "path": "csv",
            "data_files":
            ↪  ["wav2vec2_phoneme_1h/input/metadata_test_final_g2p_ipa_sample_1h.csv"],
            "text_column": "transcript_encoded",
            "path_column": "file_path"
        }
    }
}
```

Listing 1 – XLSR-APT-1h fine-tuning configuration.

```
{
    "run_name": "Wav2Vec-fine-tuning-phonemes",
    "run_description": "Fine tuning phonemes",
    "seed": 42,

    "sampling_rate": 16000,

    "num_hidden_layers": 24,

    "vocab":{
        "vocab_path": "wav2vec2_phoneme_10h/vocab.json",
        "blank": "<pad>",
        "silence": "|",
        "unk": "<unk>"
    },

    "batch_size": 32,
    "mixed_precision": true,
    "early_stop_epochs": 50,

    "epochs": 100,
    "lr": 1e-5,
    "gradient_accumulation_steps": 1,

    "logging_steps": 100,
    "load_best_model_at_end": true,
    "save_total_limit": 2,
    "warmup_ratio": 0,
    "warmup_steps": 0,
```

```
"num_loader_workers": 8,

"freeze_feature_extractor": true,
"attention_dropout": 0.1,
"activation_dropout": 0.1,
"hidden_dropout": 0.1,
"feat_proj_dropout": 0.1,
"mask_time_prob": 0.1,
"layerdrop": 0.1,
"gradient_checkpointing": true,

"output_path": "wav2vec2_phoneme_10h/output",

"dataset_cache": "wav2vec2_phoneme_10h/datasets",

"datasets":{
    "train":
        [
          {
            "name": "csv",
            "path": "csv",
            "data_files":
            ↪  ["wav2vec2_phoneme_10h/input/metadata_train_final_g2p_ipa_sample_10h.csv"],
            "text_column": "transcript_encoded",
            "path_column": "file_path"
          }
        ]
    ,
    "devel":
        [

          {
            "name": "csv",
            "path": "csv",
            "data_files":
            ↪  ["wav2vec2_phoneme_10h/input/metadata_dev_final_g2p_ipa_sample_1h.csv"],
            "text_column": "transcript_encoded",
            "path_column": "file_path"
```

```
            }
        ]
        ,
    "test":
        {
          "name": "csv",
          "path": "csv",
          "data_files":
          ↪  ["wav2vec2_phoneme_10h/input/metadata_test_final_g2p_ipa_sample_1h.csv"],
          "text_column": "transcript_encoded",
          "path_column": "file_path"
        }
  }
}
```

Listing 2 – XLSR-APT-10h fine-tuning configuration.

```
{
    "run_name": "Wav2Vec-fine-tuning-phonemes",
    "run_description": "Fine tuning phonemes",
    "seed": 42,

    "sampling_rate": 16000,

    "num_hidden_layers": 24,

    "vocab":{
        "vocab_path": "wav2vec2_phoneme_60h/vocab.json",
        "blank": "<pad>",
        "silence": "|",
        "unk": "<unk>"
    },

    "batch_size": 32,
    "mixed_precision": true,
    "early_stop_epochs": 50,

    "epochs": 100,
    "lr": 1e-5,
    "gradient_accumulation_steps": 1,
```

```
"logging_steps": 100,
"load_best_model_at_end": true,
"save_total_limit": 2,
"warmup_ratio": 0,
"warmup_steps": 0,

"num_loader_workers": 8,

"freeze_feature_extractor": true,
"attention_dropout": 0.1,
"activation_dropout": 0.1,
"hidden_dropout": 0.1,
"feat_proj_dropout": 0.1,
"mask_time_prob": 0.1,
"layerdrop": 0.1,
"gradient_checkpointing": true,

"output_path": "wav2vec2_phoneme_60h/output",

"dataset_cache": "wav2vec2_phoneme_60h/datasets",

"datasets":{
    "train":
        [
          {
            "name": "csv",
            "path": "csv",
            "data_files":
            ↪  ["wav2vec2_phoneme_60h/input/metadata_train_final_g2p_ipa_sample_60h.csv"],
            "text_column": "transcript_encoded",
            "path_column": "file_path"
          }
        ]
    ,
    "devel":
        [

          {
```

```
            "name": "csv",
            "path": "csv",
            "data_files":
            ↪  ["wav2vec2_phoneme_60h/input/metadata_dev_final_g2p_ipa_sample_1h.csv"],
            "text_column": "transcript_encoded",
            "path_column": "file_path"
          }
        ]
        ,
    "test":
          {
            "name": "csv",
            "path": "csv",
            "data_files":
            ↪  ["wav2vec2_phoneme_60h/input/metadata_test_final_g2p_ipa_sample_1h.csv"],
            "text_column": "transcript_encoded",
            "path_column": "file_path"
          }
  }
}
```

Listing 3 – XLSR-APT-60h fine-tuning configuration.