IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 September 2024, accepted 14 October 2024, date of publication 17 October 2024, date of current version 31 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3482970

RESEARCH ARTICLE

Enhancing Automatic Speech Recognition: Effects of Semantic Audio Filtering on Models Performance

YURIY PEREZHOHIN[1,2], TIAGO SANTOS[1,2], VICTOR COSTA[1,2], FERNANDO PERES[1], AND MAURO CASTELLI[2]

[1]MyNorth AI Research, 2780-125 Oeiras, Portugal
[2]NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal

Corresponding author: Yuriy Perezhohin (yperezhohin@novaims.unl.pt)

This work was supported in part by MyNorth Artificial Intelligence (AI) Research, and in part by national funds through Fundação para a Ciência e a Tecnologia(FCT) (DOI: 10.54499/UIDB/04152/2020)-Centro de Investigação em Gestão de Informação (MagIC)/NOVA Information Management School (IMS) under Project UIDB/04152/2020.

ABSTRACT This paper presents a novel methodology for enhancing Automatic Speech Recognition (ASR) performance by utilizing contrastive learning to filter synthetic audio data. We address the challenge of incorporating synthetic data into ASR training, especially in scenarios with limited real-world data or unique linguistic characteristics. The method utilizes a contrastive learning model to align representations of synthetic audio and its corresponding text transcripts, enabling the identification and removal of low-quality samples that do not align well semantically. We evaluate the methodology on a medium-resource language across two distinct datasets: a general-domain dataset and a regionally specific dataset characterized by unique pronunciation patterns. Experimental results reveal that the optimal filtering strategy depends on both model capacity and dataset characteristics. Larger models, like Whisper Large V3, particularly benefit from aggressive filtering, while smaller models may not require such stringent filtering, especially on non-normalized text. This work highlights the importance of adjusting synthetic data augmentation and filtering to specific model architectures and target domains. The proposed method, robust and adaptable, enhances ASR performance across diverse language settings. We have open-sourced the entire work, which includes 140 hours of synthetically generated Portuguese speech, as well as the pipeline and parameter settings used to create these samples. Additionally, we provide the fine-tuned Whisper models and the code required to reproduce this research. Our code will be available at https://github.com/my-north-ai/semantic_audio_filtering.

INDEX TERMS Automatic speech recognition, contrastive learning, data augmentation, embeddings, synthetic data filtering, text-to-speech.

I. INTRODUCTION

In recent years, the field of Automatic Speech Recognition (ASR) has witnessed remarkable progress, thank to the combination of diverse datasets and innovative modeling techniques. Central to this advancement is the integration of both natural and synthetic speech data [1], [2], [3]. While natural speech data provides a rich representation of real-world acoustic and linguistic diversity, synthetic data generated through Text-to-Speech (TTS) models offer a scalable and cost-effective way to augment training corpora, particularly for low-resource languages or specialized domains. However, the effectiveness of synthetic data depends on its quality and relevance to the target task [4], [5]. Thus, it is necessary to use robust filtering methods to ensure that only high-quality samples are incorporated into the training process.

In this paper, we propose a methodology for filtering synthetic audio data based on contrastive learning [6]. By aligning representations of spoken language and corresponding text transcripts, our approach aims to identify

The associate editor coordinating the review of this manuscript and approving it for publication was Lorenzo Mucchi.

and remove synthetic samples that exhibit poor semantic or acoustic correspondence. This not only improves the overall quality of the training data but also enhances the ASR model's ability to generalize to unseen speech patterns. Additionally, we implement a technique to ensure data integrity on all the training sets, which is based on words spoken per second.

The main contributions can be summarized as follows:

- We propose a novel methodology for filtering synthetic audio data, ensuring the inclusion of only high-quality samples for ASR training.
- We investigate the impact of different filtering thresholds and model sizes on ASR performance, providing insights into the optimal strategies for using synthetic data.
- We demonstrate significant improvements (in terms of word error rate) for Portuguese ASR by incorporating filtered synthetic data, showing the effectiveness of the proposed method in a medium-resource language setting.

The structure of this paper is organized as follows: Section II provides a detailed overview of related work in the areas of ASR, audio representation learning, and contrastive learning. Section III describes the proposed methodology, including data collection and the contrastive learning framework for filtering synthetic data. Section IV presents the experimental setup, including datasets, model configurations, and evaluation metrics. Section V reports the results of the experimental phase, highlighting the impact of the proposed methodology on ASR performance. Finally, Section VI concludes the paper discussing the main findings of this work and suggesting directions for future work.

II. RELATED WORK

This section provides a complete overview of the current state of Automatic Speech Recognition (ASR), highlighting its evolution from classical methods to deep learning techniques. We examine works that use multiple data sources, including crowd-sourced datasets [7], [8], multilingual corpora, and synthetically generated audio [1], [2], [3], to train ASR models.

Additionally, we conduct a thorough examination of existing work on embeddings, both text and audio. We analyze the limitations of traditional hand-crafted audio features and the subsequent shift towards self-supervised learning (SSL) for audio representation. We then explore recent advancements in SSL for audio, with a particular focus on contrastive learning techniques, a central element of the proposed methodology. Finally, we investigate the field of multimodal contrastive learning, examining how it can be applied to learn joint representations of audio and text.

A. AUTOMATIC SPEECH RECOGNITION

1) CLASSICAL APPROACHES

Transforming audio signals carrying speech information into plain text has been a long line of research, beginning with the introduction of the Automatic Digit Recognizer *Audrey* back in 1952 [9]. The algorithm behind this machine recognized stored energy patterns and produced a probability for each digit from zero to nine. The authors noted that *Audrey* required more energy patterns to classify correctly, as the distribution of energy signals varied significantly between speakers. Following *Audrey*, the *Shoebox* algorithm developed in 1962 was capable of recognizing 16 spoken words. The algorithm converted audio signals into electrical impulses, which were then classified into different sound types by a measuring circuit. Between 1971 and 1976, ASR gained more popularity with the development of a new system named *Harpy* [10], which could understand over 1000 words. The *Harpy* system performed parametric analysis and segmentation on input speech and matched speech segments with predefined templates using the Itakura distance [11] to identify spoken words. Subsequently, Beam Search was employed to efficiently navigate through a vast network of possible word sequences represented as a graph of phonetic and word components. This technique allowed *Harpy* to focus on the most promising paths, reducing computational resources and complexity. These foundational systems paved the way for continued advancements in methods for transforming audio signals into text.

Predominant statistical ASR systems used several blocks to transform an audio signal into a human-readable transcript [12]. Initially, the signal was processed to reduce noise and distortions, and then features were extracted using methods such as mel-frequency cepstral coefficient (MFCC) [13] or relative spectral transform-perceptual linear prediction (RASTA-PLP) [14]. MFCC captures the audio spectrum on a nonlinear scale, providing a compact and efficient representation of perceptually important features. In contrast, RASTA-PLP applies perceptual models to filter out slow and fast variations in the speech signal, focusing on critical filters for speech perception. Following the features extraction, acoustic models, such as Hidden-Markov Models [15] (HMM) or Gaussian Mixture Models [16](GMM), are trained by estimating the maximum likelihood parameters from the extracted features. Finally, the pronunciation model, typically constructed as a set of rules by a human linguist, maps phonemes into letters, and a language model converts the sequence of characters into the final transcript.

These models enhance the performance of ASR systems by learning the statistical properties of speech signals, making them more effective than earlier memory-based approaches like *Audrey* or *Shoebox*. They combine statistical methods and probability theory to model complex probabilistic relationships in speech. For instance, some authors have extended estimation techniques to work with Gaussian distributions [17]. The system iteratively improves performance by breaking down speech signals into smaller parts and estimating parameters using HMMs and GMMs, allowing for more accurate and robust speech recognition. Another method that improved the existing implementations of HMMs and GMMs involves changing the parameter estimation technique

[18]. The method focuses on maximizing mutual information between the signals and the word sequence rather than using maximum likelihood estimation. This technique enhances model discrimination and improves word recognition accuracy compared to the previous implementations.

While GMMs and HMMs have improved sequential speech recognition, their linearity assumption and limitations in modeling long-term dependencies [19] are significant constraints, as they can not model non-linear functions. These drawbacks pushed researchers to focus on more robust and innovative methodologies, such as Deep Neural Networks (DNN) [20]. These networks were initially used to extract optimal or near-optimal features, enriching the input to the GMM and HMM models [21].

2) DEEP LEARNING APPROACHES

The creation of DNNs allowed researchers to develop models with human-like capabilities. However, the rise of these models only occurred in the last decade due to prior constraints on publicly available data and computational resources. Initially, DNNs contributed to ASR systems by extracting signal features for further processing.

DNNs were used in tandem with traditional GMMs and HMMs, providing discriminative capabilities for feature extraction that improved the accuracy of ASR systems [22]. By generating posterior probabilities for subword units, these DNNs enabled more precise acoustic modeling. This combination led to substantial error rate reductions, as demonstrated in noisy environments like the Aurora [23] task, where these systems achieved a relative error rate reduction of over 35%. The advanced feature processing capabilities allowed for more accurate and robust speech recognition, improving ASR performance by better capturing the complexities of spoken language. A few years later, researchers explored using a multi-layer perceptron (MLP) with a bottleneck layer to produce input features directly for the GMM-HMM recognition system, eliminating the need for separate probabilistic feature extraction [21]. This approach further enhanced ASR accuracy and efficiency by providing a more informative representation of speech signals [24].

Moreover, with the increased use of Convolutional Neural Networks [25] (CNN) in computer vision, some researchers applied them to speech recognition tasks [26]. CNNs are well known for their capabilities to extract features from images using local filtering and pooling techniques. One study [27] proposed using local filtering and max-pooling to normalize speaker variance in combination with HMM, where the latter received state probabilities generated by CNNs. In the final phase, a Viterbi [28] decoder produced the word labels corresponding to the input speech sequence. The authors concluded that using convolutional layers with 84 filters and a pooling size of 6 significantly reduced errors in the *TIMIT* dataset [29].

ASR models have significantly improved with the advent of DNNs. Instead of constructing complete systems with different blocks, researchers began focusing on the end-to-end neural models that address the task from feature extraction to word prediction from audio inputs [30].

One of these end-to-end models appeared with the combination of CNNs and DNNs, which were jointly trained [31]. The objective was to develop a model that merges features from both networks by concatenating their outputs and passing them through subsequent dense layers for prediction. Additionally, the authors investigated using features-space Maximum Likelihood linear regression (fMLLR) to warp the log-mel features for the input, along with different pooling, dropout, and weight-sharing strategies. Recurrent Neural Networks [32] (RNN) followed the progress in speech recognition and achieved major success due to their ability to model time sequences, such as speech signals. Some researchers [33] implemented RNNs as a solution for hybrid systems, which did not fully exploit the potential of RNNs for sequence modeling. The research introduced a method of using RNNs for labeling sequence data, eliminating the need for pre-segmented training and post-processed output. However, RNNs face a relevant drawback in modeling long-term dependencies due to the vanishing gradients problem, where gradients diminish exponentially during backpropagation, making it difficult to learn and retain information over long sequences. To overcome the vanishing gradient problem associated with RNNs, researchers adopted the Long Short-term Memory [34] (LSTM) architecture, specifically designed to maintain long-term dependencies in sequence data. The authors demonstrated significant improvements using LSTMs in speech recognition accuracy on the TIMIT benchmark [35].

With the introduction of transformers [36], the ASR task gained even more popularity as the architectures and the dimensionality of available data became more robust. The Transformer architecture relies on self-attention layers, making it easier to capture extensive dependencies in input speech, consequently outperforming previous models and systems [37]. The *Whisper* [38] architecture is one of the best-performing models across several benchmarks in diverse languages, such as CommonVoice and MLS. *Whisper* employs an encoder-decoder Transformer architecture, where the encoder processes log-mel spectrogram representations of the audio and the decoder predicts the transcript tokens. This model is trained on a vast dataset of 680000 hours of multilingual and multitask audios, making it robust for generalization across various languages. *Whisper* has made significant progress in ASR, with many researchers leveraging the pre-trained model and fine-tuning for domain-specific languages and dialects [39], [40], [41]. Besides *Whisper*, other models have been developed, such as *SeamlessMT4* [42], which can transform text into speech waveforms in different languages, and *Speecht5* [43], a successor of the original *T5* text-to-text generation model [44].

Despite the capabilities of the large and robust models, considerable research is still needed to address challenges with low-resource languages. One study explored different

methodologies and fine-tuning strategies for *Whisper* on seven low-resource languages [45]. The authors pointed out issues in the ASR task when using the *Whisper* model for languages that are uncommon or lack abundant data. Due to these problems and ongoing innovation in the natural language processing field (NLP), new ideas have emerged, such as combining text-to-speech models for data augmentation to enhance the fine-tuning process of ASR models [46]. One study fine-tuned *Whisper* only on synthetic data for low-resource languages [47], achieving word error rate reductions between 2 and 30 points. Another research focused on augmenting audio-text pairs with synthetic speech for four minority languages: West Germanic, Gronings, West-Frisian, Malayo-Polynesian, Besemah, and Nasal [1]. The authors used pre-trained models to generate synthetic transcriptions for humanly transcribed audio and employed a fine-tuned text-to-speech (TTS) model for the Gronings language, comparing performance across all experiments. They reported a significant performance increase with their self-training method and an even greater word error rate reduction for the ASR model augmented with TTS. However, they posed an open question about whether synthetically augmented data is always beneficial for larger datasets of highly curated audio-transcription pairs. Using data augmentation can have drawbacks, such as mismatches between the real and generated speech, which can lead to limited improvement or even decreased performance [48]. A plausible solution proposed training a neural network to measure the similarity of synthetic samples to real speech [4]. The authors discovered that including samples with considerable dissimilarity can improve the ASR model's performance by introducing more lexical differences. However, samples produced by the TTS that are too dissimilar can degrade the recognition performance, raising questions about effectively selecting synthetic data.

B. EMBEDDINGS

In NLP, extracting meaningful information from real-world data such as text and audio requires converting these data into a format that computers can understand and process. Embeddings are a fundamental tool in this transformation, providing a way to represent complex objects as dense vectors within a continuous, multi-dimensional space. Each dimension of this space captures a specific feature or attribute of the object, whether it be semantic meaning, syntactic structure, or acoustic properties in the case of audio.

The power of embeddings lies in their ability to map semantic or acoustic similarity between real-world objects to spatial proximity in the embedding space. By representing similar objects with vectors that are close together, machine learning models can easily identify and leverage these relationships for a wide range of tasks, such as semantic search, question answering, or machine translation.

1) WORD EMBEDDINGS

Early word representations, such as BoW [49], were limited in their ability to capture the nuanced semantic relationships between words due to their reliance on treating words as independent entities, ignoring the role of context in shaping meaning.

These limitations motivated a shift towards neural network approaches, which marked a turning point in the field of word embeddings. Models like Word2Vec [50] and GloVe [51] emerged, leveraging neural networks to learn embeddings by considering the distributional properties of words - essentially, predicting a word based on its surrounding context or vice versa. This shift enabled the capture of both semantic relationships and distributional semantics, leading to significant advancements in natural language understanding.

Building upon this foundation, the emergence of RNN based models and attention-based mechanisms [52] led to the development of contextual embeddings. However, it was the introduction of the Transformer architecture [53], that truly revolutionized the field.

The Transformer's encoder component, designed to extract meaningful representations from textual data, has given rise to powerful text encoders, with BERT [54] being the most prominent example. BERT's bidirectional training and pre-training on massive datasets have made it particularly effective at capturing nuanced linguistic patterns that were previously challenging for earlier models.

BERT has since become the foundation of numerous state-of-the-art models in text encoding, each with its own unique strengths. These include RoBERTa [55], which optimized BERT's training methodology. ALBERT [56], a lighter and more efficient variant, and DeBERTa [57], which incorporates disentangled attention mechanisms. Beyond these, models like Sentence-BERT [58], a modification of BERT using siamese and triplet networks, have specifically addressed the challenges of sentence-level embeddings, finding applications in tasks like semantic search.

2) AUDIO EMBEDDINGS

The real world is filled with sounds, from the human voice to the elaborate melodies of music. This data is rich in information, but in its natural state, it exists as continuous waves of pressure traveling through the air. To analyze this data with computers, it must be converted into a digital format. Microphones capture these waves and convert them into analog electrical signals, which are then transformed into discrete digital representations through sampling and quantization. Sampling involves measuring the amplitude of the analog signal at regular intervals dictated by the sampling rate. Quantization then maps these continuous amplitude values to a finite set of discrete levels. These two steps produce a digital audio signal, represented as a waveform, a series of discrete amplitude values recorded at specific time

intervals. This waveform serves as the raw input for further analysis and feature extraction.

From this raw waveform, we can extract features that capture different aspects of the audio signal:

- **Time-domain features**: derived directly from the waveform, these features describe the temporal characteristics of the sound:
  - Zero-Crossing Rate (ZCR): the rate at which the waveform crosses the zero-amplitude line, often used to distinguish voiced and unvoiced segments in speech. [59] employed ZCR, Spectral Centroid, and other features in conjunction with AdaBoost for music genre classification
  - Amplitude Envelope: the overall contour of the waveform's amplitude, providing insights into the energy dynamics of the sound.
- **Frequency-domain features**: obtained by transforming the waveform into the frequency domain, typically using the Fourier Transform, these features reveal the distribution of energy across different frequencies:
  - Spectral Centroid: the average frequency weighted by amplitude, providing a measure of the sound's brightness.
  - Peak Frequency: the frequency with the highest amplitude, often representing the dominant pitch in a sound.
- **Cepstral-domain features**: derived from a further transformation of the frequency domain. First, the logarithm of the magnitude spectrum is calculated, then an inverse Fourier Transform is applied:
  - Mel-Frequency Cepstral Coefficients: the most widely used cepstral features, MFCCs are calculated using a Mel-scale filter bank applied to the power spectrum, mimicking how humans perceive sound
  - Linear Prediction Cepstral Coefficients (LPCCs): an alternative to MFCCs, these coefficients are based on linear prediction of the spectral envelope. [60] utilized LPCCs for noise removal, while MFCCs have been leveraged for ASR and speech enhancement [61], [62]

Traditional machine learning approaches [59], [60], [61], [62] rely on manually engineered features extracted from these domains. However, the abundance of potential audio features makes manual selection a challenge. Moreover, these hand-crafted features may not generalize well across different audio tasks and domains.

The limitations of manual feature engineering, combined with the increasing availability of large unlabeled audio datasets, have driven the exploration of alternative methods, particularly Self-Supervised Learning (SSL), which have the potential to learn relevant features directly from raw audio data.

Early SSL methods for audio include CPC [63], which trains models to predict future audio segments based on past context. This approach was further revolutionized by Wav2Vec 2.0 [64], where portions of the raw audio waveform are masked, and the model is trained to predict the masked portions based on the surrounding context. HuBERT [65] builds on this concept with an additional clustering step for more efficient and discriminative representations. WavLM [66] incorporates both masked speech prediction and denoising objectives. The ability of these models to generate high-quality audio embeddings has proven highly effective, not only in core tasks like ASR but also in a wider range of applications, including emotion recognition and music information retrieval.

C. CONTRASTIVE LEARNING

The abundance of unlabeled data, which is often expensive and time-consuming to label, has motivated the development of SSL. SSL represents a paradigm shift in machine learning by enabling models to learn meaningful representations directly from unlabeled data, thus bypassing the need for extensive human annotation. This is achieved through pretext tasks that exploit inherent data structures, serving as pseudo-supervision to guide the model in learning representations that are useful for downstream tasks.

Among the various pretext tasks, Contrastive Learning (CL) has emerged as a dominant and highly effective approach in SSL. CL is fundamentally a discriminative technique that aims to group similar samples closer together while pushing dissimilar ones farther apart. This is achieved by training models on positive and negative pairs, where positive pairs are typically created through data augmentation techniques applied to the same sample, and negative pairs are different samples.

CL models learn to distinguish between positive and negative pairs by measuring the similarity between their embeddings in a latent space, often derived from features extracted by an encoder network. This process is achieved with a variety of loss functions such as triplet loss [67], InfoNCE [63], N-pair [68], and NT-Xent [69]. These losses are designed to encourage the model to learn representations that are invariant to irrelevant transformations, leading to the creation of high-quality representations that are transferable to downstream tasks.

1) UNIMODAL CONTRASTIVE LEARNING

Unimodal CL is a cornerstone in the evolution of representation learning, providing a robust foundation for understanding individual modalities such as text, images, and audio. By leveraging the inherent structure and patterns within each modality, unimodal CL methods can learn meaningful representations without relying on external labels, thus addressing the challenge of limited labeled data.

In NLP, Translation Language Modeling (TLM) extends masked language modeling (MLM) [54] by operating on parallel sentences in different languages, training a model to predict masked words in one language by leveraging

information from both. This approach enhances cross-lingual transferability and improves performance on tasks such as cross-lingual classification and question answering.

In the field of time series analysis, contrastive based methods like Temporal Neighborhood Coding (TNC) [70] have shown promise in learning effective representations of sequential data. By contrasting neighboring and non-neighboring signals, TNC can capture temporal dependencies and enhance performance in clustering and classification tasks.

Computer vision has also witnessed significant advancements, especially with the Deep InfoMax (DIM) [71] approach, which maximizes mutual information between representations extracted from multiple views of an image, effectively learning discriminative features. SimCLR [69] further extends this concept by contrasting augmented views of the same image with those of different images, leading to robust image representations that excel in downstream tasks.

The audio domain has similarly benefited from unimodal SSL, with models like [63], [64], and [65] achieving impressive results in speech recognition tasks. These models leverage various self-supervised objectives, such as contrastive predictive coding and masked prediction, to learn representations that capture the complex structure and semantics of audio signals. CLAR [72] applies a contrastive objective to both raw audio and mel-spectrograms, while COLA [73] focuses on mel-spectrograms with bilinear comparisons. Additionally, SoundSemantics [74] employs Siamese networks and Euclidean distance to determine if audio pairs belong to the same class, and [75] explores multi-format learning using Siamese networks and contrastive learning for robust audio representations.

2) MULTIMODAL CONTRASTIVE LEARNING

While unimodal approaches focus on individual modalities such as text and audio, our experience of the world is inherently multimodal, consisting of a rich interplay of different information sources. This has led to an increased interest in multimodal contrastive learning, aiming to bridge the gap between modalities and learn joint representations.

Although the majority of research in multimodal representation learning has traditionally focused on the image-text modality [76], [77], [78], driven by the relative abundance of labeled image-caption data [79], [80], [81], [82], the field of audio-text integration is rapidly growing with availability of large-scale labeled audio-text datasets, such as [83], [84], and [85].

Both SoundSemantics [74] and SLAM [86] enhance audio representations by incorporating textual information. SoundSemantics integrates pre-trained word embeddings from Word2Vec [50] into an acoustic event classification model, while SLAM jointly pre-trains a unified encoder on text and speech data using BERT and wav2vec 2.0-based objectives for downstream tasks like ASR and speech translation.

CLIP [87], an influential model in the image-text domain, that comprises two main components: a text encoder and an image encoder. The model is designed to learn by maximizing the cosine similarity between matching text and image embeddings while minimizing it for mismatched pairs, effectively establishing a joint understanding of both modalities. Clip has inspired extensions to audio-text modalities like AudioCLIP [88], Wav2CLIP [89], and CLAP [90] are multimodal models that build upon or are inspired by CLIP, utilizing contrastive learning to align audio representations with text and/or images. AudioCLIP and CLAP use a dual-encoder architecture, while Wav2CLIP employs a two-stage approach with pre-training on visual information followed by fine-tuning on audio-text pairs.

Most recently, [91] introduced a pipeline for contrastive language-audio pretraining that further enhances audio representations by combining audio data with natural language descriptions, utilizing feature fusion and keyword-to-caption augmentation to improve the model's handling of varying audio lengths and enrich textual context.

Beyond pairwise modalities, research has extended to the fusion of image, audio, and text. Both VATT [92] and [93] frameworks leverage contrastive learning and self-supervision to learn multimodal audio representations, combining audio with video data for enhanced feature learning. While VATT explores different architectures for video-audio-text integration, [93] work focuses on using video to improve general audio representations.

More recently, ONE-PEACE [94] has emerged as a highly extensible model designed to unify representation learning across vision, audio, and language modalities through two modality-agnostic pretraining tasks: cross-modal contrastive learning for aligning semantic spaces between modalities, and intra-modal denoising contrastive learning for learning robust representations within each modality.

III. METHODOLOGY

This section describes all the steps of the proposed methodology. We first formally define the two primary data types utilized in this work: audio and textual data. The Data Collection subsection describes the curated Portuguese datasets and the generation of synthetic speech using the TTS model. In Contrastive Learning-Based Synthetic Data Filtering, we introduce the filtering methodology, focusing on improving synthetic data quality. The subsequent subsection covers the final step in audio processing, which relies on speech rate to ensure dataset integrity. Lastly, Model Refinement and Evaluation explains the fine-tuning process. Figure 2 provides a detailed overview of the methodology developed in this study.

A. AUDIO REPRESENTATION

As discussed in Section II-B2, raw audio signals are continuous waveforms that undergo sampling and quantization to enable computational analysis. We define a digital audio

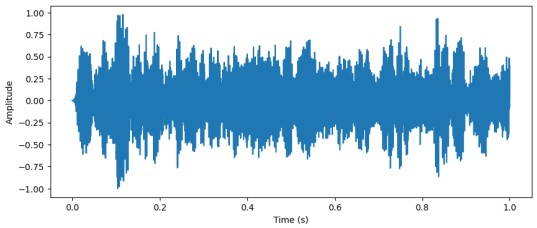

FIGURE 1. Illustration of a synthetic waveform.

signal in the time domain as a discrete sequence of N samples:

$$x[i] = (t_i, a_i), \quad i = 0, 1, 2, \ldots, N-1 \tag{1}$$

where:

- $t_i \in \mathbb{R}$ represents the time instant of the i-th sample, determined by the sampling rate f_s (typically measured in Hertz), and given by:

$$t_i = \frac{i}{f_s} \tag{2}$$

- $a_i \in \mathbb{R}$ represents the quantized amplitude of the signal at time instant t_i.

This digital representation, often visualized as a plot of amplitude (a_i) versus time (t_i), is known as a raw waveform (Figure 1).

B. TEXTUAL DATA REPRESENTATION

Textual data, inherently discrete in nature, consists of sequences of distinct symbols (characters, words, or subword units). Similar to audio, however, raw text requires transformation into numerical representations to be processed by ML models. When using Transformers [53], this is typically achieved through tokenization, where a sentence is divided into its constituent words or subword units.

$$S = (t_1, t_2, \ldots, t_M), \quad i = 0, 1, 2, \ldots, M-1 \tag{3}$$

where:

- t_i represents the i-th token in the sentence.
- M is the total number of tokens in the sentence.

The choice of tokenization scheme depends on the specific model and task. Common approaches include word-level tokenization, subword tokenization (e.g., Byte Pair Encoding), and character-level tokenization.

C. DATA COLLECTION

Our initial goal was to enhance ASR performance in Portuguese by adopting a fine-tuning approach on the Whisper model. This approach builds upon insights from [95] that highlight the correlation between model performance and training data volume. We curated three Portuguese datasets for this purpose: the training subset of MLS [8], the training subset of Common Voice 16.1 [7], and Perfil Sociolinguístico da Fala Bracarense (PSFB) [96], the latter being fully Portuguese dialect.

Following previous research demonstrating the benefits of synthetic data for ASR improvement [1], [2], [3], we utilized a state-of-the-art TTS model to generate synthetic speech. Further details on the specific datasets used, both real and synthetic, and their respective sizes will be provided in Section IV.

D. CONTRASTIVE LEARNING-BASED SYNTHETIC DATA FILTERING

Initial fine-tuning of a Whisper Small model on the combined real and synthetic data yielded unsatisfactory results. Thus, we explored methods for enhancing the quality of the training data.

Recognizing that the quality of synthetic audio data significantly impacts ASR performance, we developed a novel filtering methodology inspired by MusCALL [97]. This approach uses contrastive learning to train a model that encodes both audio and text into a shared multidimensional space, aiming to maximize the similarity between corresponding audio-text pairs.

The goal of the model is to learn projection layers, denoted as $g_a(\cdot)$ for the audio modality and $g_t(\cdot)$ for the text modality, that map the pre-trained embeddings from the encoders into a shared multimodal embedding space. These projection layers are essential for aligning the representations of different modalities and enabling effective contrastive learning.

As training data, we used paired data consisting of text transcriptions and their corresponding audio waveforms, extracted from the Portuguese train subset of Common Voice 16.1 a more robust and and up-to-date dataset compared to MLS, as the primary source of training data.

Each transcription T_i is represented as a sequence of tokens $T_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,M})$, where $t_{i,j}$ is the j-th token, and M is the number of tokens in the i-th transcription. Each audio waveform A_i is represented as a sequence of samples $A_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,N})$, where $a_{i,j}$ represents the amplitude at the j-th time step, and N is the number of samples in the i-th waveform. For text encoding, we utilize a DeBERTa-based model [98] pretrained on Portuguese data, which transforms each transcription T_i into a text embedding z_{T_i}. For audio encoding, we employ Whisper Medium, which encodes each audio waveform A_i into an audio embedding z_{A_i}.

The pre-trained encoders are frozen during training, and the focus is on optimizing the parameters of the projection layers. The objective is to learn these projection layers such that the resulting multimodal embeddings, $g_a(z_{A_i})$ and $g_t(z_{T_i})$, are close in the shared space for matching audio-text pairs, while mismatched pairs are pushed apart.

In implementing the training process of contrastive learning, we draw upon the insights presented in [97], which highlights potential limitations in the traditional approach of constructing positive and negative pairs through instance discrimination. Specifically, traditional approaches assume that all items inside of a dataset are perfectly aligned, with each audio-text pair representing the most semantically

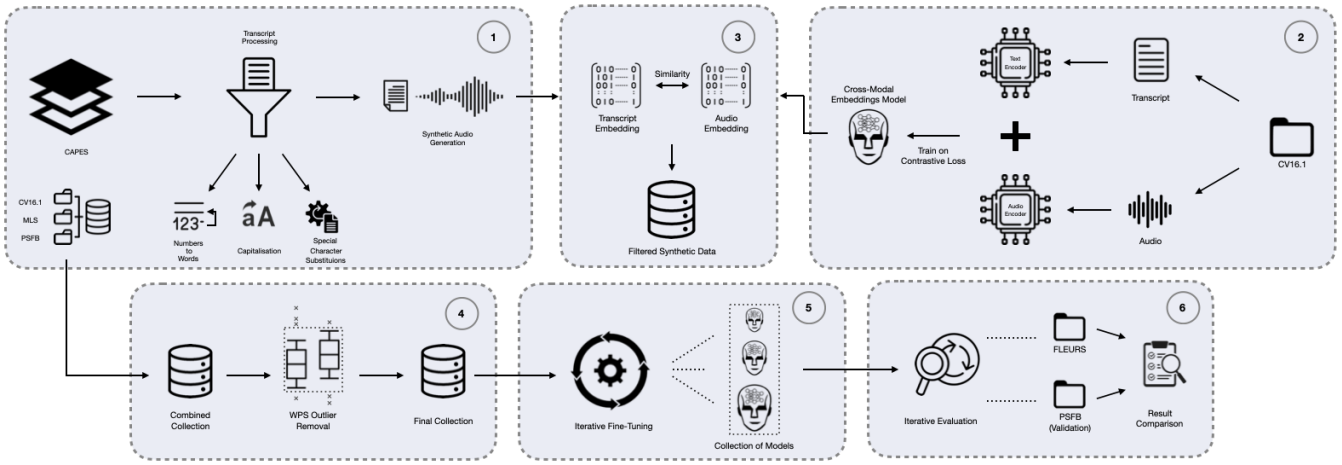

FIGURE 2. **Detailed methodology overview: The process begins with the collection and preprocessing of transcripts for synthetic audio generation. The second step employs a contrastive learning approach for the cross-modal embedding model. In the third step, this model calculates the similarity between the synthetic audio and corresponding transcripts, creating several subsets of filtered data. The fourth step involves final preprocessing based on words per second (WPS) for the combined audio collection, which includes both synthetic and real data. The fifth step outlines the iterative fine-tuning process applied to each data subset (filtered and unfiltered). Finally, the last step demonstrates the evaluation of each model on two distinct data subsets.**

relevant match, and that all non-aligned pairs are equally dissimilar. However, in real-world datasets within a randomly sampled mini-batch, some samples will share similarities with other tracks, and so will their respective captions often exhibit varying degrees of similarity between items, both within and across modalities.

To address this, we adopted a refined sampling strategy [97], which leverages the observation that similar captions often correspond to similar audio content. This strategy estimates the relevance between non-paired (negative) items in a mini-batch and assigns weights to negative samples based on their relevance to the anchor sample. This relevance-based weighting scheme is implemented through the following formula:

$$w_i = \exp\left(\frac{\frac{1}{N}\sum_{j=1}^{N} \text{sim}(T_i, T_j)}{\kappa}\right) \tag{4}$$

where:

- $\text{sim}(v_i, v_j)$ is the similarity score between the text embeddings of the anchor sample v_i and the negative sample v_j. This similarity score is computed using a pre-trained sentence transformer model[1].
- κ is a temperature hyperparameter that controls the sharpness of the weighting distribution.
- N is the batch size.

By incorporating this relevance-based weighting, the audio-to-text contrastive loss *Eq.*5 is modified inspired by the InfoNCE loss [63], which takes into account the varying relevance of negative samples. This loss function can be denoted as:

$$\mathcal{L}_{a\to t} = -\frac{1}{N}\sum_{i}^{N} w_i \log \frac{\exp(z_{a,i}\cdot z^{+}_{t,i}/\tau)}{\sum_{z\in\{z^{+}_{t},z^{-}_{t}\}}\exp(z_{a,i}\cdot z/\tau)} \tag{5}$$

[1]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

where:

- $z^{+}_{t,i}$ are embeddings of positive text samples for A_i.
- $z^{-}_{t,i}$ are embeddings of negative text samples for A_i.
- τ is a temperature hyperparameter to scale similarity scores.
- w_i is the relevance-based weighting of the i-th sample.
- N is the batch size.
-

By noting that $\mathcal{L}_{t\to a}$ can be defined symmetrically to Eq.(1), the total loss over all (audio, text) pairs in a mini-batch is simply obtained by summing the two losses together:

$$L_{a,t} = \mathcal{L}_{a\to t} + \mathcal{L}_{t\to a} \tag{6}$$

Unlike typical applications of such models for retrieval or zero-shot transfer, we utilized the learned representations for an innovative downstream task: quantifying the similarity between synthetic audio and its corresponding text transcript. We used cosine similarity as the metric and experimented with various thresholds, set at 1, 2, and 3 standard deviations below the mean cosine similarity between audio and text embeddings (Table 1).

TABLE 1. **Synthetic data filtering setup.**

SD	Utterances Removed (%)	Similarity Threshold
1	17.68	0.68
2	3.92	0.64
3	0.24	0.60

This filtering process generated three distinct synthetic datasets of different sizes, each reflecting a different similarity threshold. These filtered datasets were then combined with three real speech datasets (MLS, CV, and PSFB) to create three augmented training sets. Additionally, we constructed a dataset comprising only the real speech data for comparison.

E. POST-FILTERING: REMOVING OUTLIERS

Following the filtering of synthetic data and its integration with the real speech datasets, a thorough examination of the resulting datasets was conducted. This assessment revealed the presence of real speech samples with atypical speech rates. Some samples exhibited abnormally slow speech rates, often accompanied by transcription errors where the provided text did not accurately reflect the spoken content. Other samples displayed unusually high speech rates, indicating rapid speech that often coincided with degraded audio quality.

To address this issue and ensure the integrity of the training data, we implemented a final filtering step based on words per second (WPS), as illustrated in Figure 3.

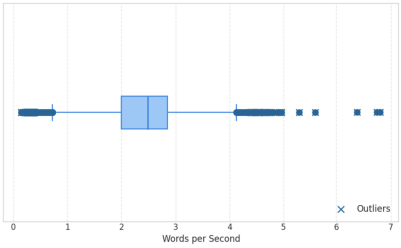

FIGURE 3. Distribution of words per second with outliers (using 3σ).

We calculated the WPS for each utterance in the combined datasets and identified outliers that deviated significantly from the typical speech rate, which is a strong indicator of transcription errors. We chose a threshold of three standard deviations from the mean WPS to define outliers. This conservative threshold allowed us to remove only the most extreme cases, preserving the majority of the data while excluding potentially detrimental samples.

The impact of this process is shown in Table 2, where slight variations in the number of utterances for the real speech datasets (MLS, CV, and PSFB) are evident across different filtering thresholds.

F. MODEL REFINEMENT AND EVALUATION

With the filtered synthetic and original data, we fine-tuned a collection of Whisper models (small, medium, and large) on varying combinations of real and synthetic data. By integrating established techniques with novel methodology, the proposed approach demonstrates the potential of contrastive learning for synthetic data filtering and its impact on enhancing ASR model performance.

IV. EXPERIMENTAL SETUP

A. DATASET

In this subsection, we outline the composition of the training and validation datasets.

- **Real Audio Data:** We utilized three Portuguese datasets as source of real audio data:
  - Multilingual LibriSpeech Corpus (MLS) [8]: We used the Portuguese train subset of MLS, a widely-used open-source dataset.
  - Common Voice (CV) 16.1 [7]: Another publicly available and widely recognized dataset of transcribed speech. We utilized the Portuguese train subset.
  - Perfil Sociolinguístico da Fala Bracarense (PSFB) [96]: This regional dataset captures the distinct speech patterns of Braga, Portugal, including the prevalent phenomenon of "queismo" (omission of prepositions), which is observed in 84% of utterances. This unique linguistic characteristic, united with the Alto Minho accent, makes PSFB a challenging dataset for ASR models. We utilized speech from this corpus to assess the methodology's effectiveness on non-standard Portuguese speech patterns.
- **Synthetic Audio Data:** Previous research has shown that training ASR models exclusively with synthetic data can still produce significant results with minimal overfitting [46]. In this research, we generated 48972 utterances of synthetic data, approximately 140 hours, using the SeamlessM4T-v2 model [99]. The text data for TTS was sourced from the CAPES dataset [100], a collection of theses and dissertations from various knowledge areas. Preprocessing was applied to remove numbers, special characters, and mathematical equations. Also, capitalization was applied before TTS generation. To ensure diversity, we created three synthetic speakers (two male and one female).

1) TRAINING SETS

To investigate the influence of synthetic data quality, we constructed four training datasets:

- **Zero Synthetic:** Only real audio data (MLS, CV, partial PSFB).
- **Filtered Synthetic (FS-X%):** Real data combined with varying amounts of synthetic data filtered (X% removed) using the contrastive learning-based methodology (detailed in Section III-D).

2) VALIDATION AND TEST SETS

For validation, we created a dataset with a similar distribution to the training sets, using the test subsets of MLS and CV, along with a split from PSFB.

Detailed statistics on the duration and number of utterances for each dataset are summarized in Table 2.

To establish a rigorous evaluation of our approach, we evaluated the models using two distinct sets of data. The first is the Portuguese test subset of the FLEURS dataset [101], an independent dataset that allows us to assess performance on unseen data and provides a broader evaluation of the models' capabilities. The second is the PSFB subset present in the validation set, chosen due to its unique linguistic and acoustic challenges, including the prevalence of specific Portuguese dialect.

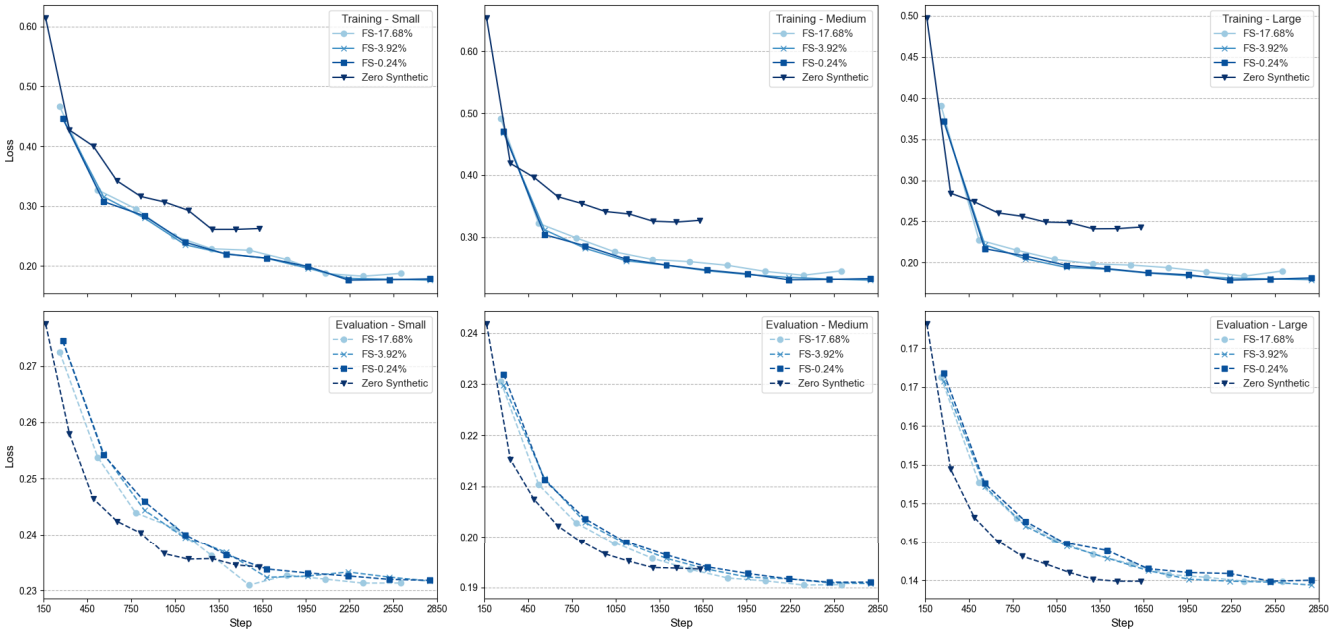

FIGURE 4. Training and evaluation losses of whisper models fine-tuned on filtered synthetic data FS settings and zero-synthetic. The top row represents the training loss curves. The bottom row shows the corresponding evaluation loss curves. Due to the smaller dataset size, the zero-synthetic configuration completed fine-tuning earlier. The FS-0.24% configuration, with the most synthetic data retained, required the longest training time. Refer to Table 2 for details on the composition of each dataset.

TABLE 2. Composition of training and validation datasets after synthetic filtering and outlier removal.

	Zero Synthetic	FS-17.68%	FS-3.92%	FS-0.24%	Validation
Duration (h)					
Synthetic	0.00	101.50	129.83	139.86	0.00
MLS	160.96	160.89	160.87	160.87	3.64
CV	24.58	24.34	24.11	24.06	11.53
PSFB	62.43	63.17	63.18	63.20	6.66
Total	247.97	349.90	378.00	387.99	21.83
Nr. Utterances					
Synthetic	0	40,314	47,048	48,852	0
MLS	37,532	37,516	37,512	37,511	826
CV	21,525	21,265	20,975	20,906	9,362
PSFB	10,462	10,587	10,589	10,591	1,116
Total	69,519	109,682	116,124	117,860	11,304

B. EVALUATION METRICS

To assess the effectiveness of the proposed methodology, we utilize one key metrics:

- **Word Error Rate (WER):** WER quantifies the dissimilarity between a predicted transcript and the reference (ground truth) transcript at the word level. It is calculated as the sum of substitution (S), deletion (D), and insertion (I) errors divided by the total number of words (N) in the reference:

$$WER = \frac{S + D + I}{N} \tag{7}$$

C. CONTRASTIVE LEARNING MODEL TRAINING

Following MusCALL [97] for the loss scaling weight w_i, we use the cosine similarity between L2-normalized embeddings from the pre-trained sentence transformer model. In Equation 4, we set the loss weighting temperature parameter κ to 0.01. Moreover, instead of manually tuning the temperature parameter τ in Equation 5, we incorporated it into the model's learning process.

We trained the model for 60 epochs on the Portuguese train subset of Common Voice 16.1, using a batch size of 32 and an initial learning rate of 3e−5. We utilize the AdamW optimizer with a cosine annealing learning rate scheduler for efficient training. The computational resources for training were provided by a g5.xlarge instance on AWS. After the training process, we selected the best model based on its performance on the validation loss.

D. WHISPER FINE-TUNING

All Whisper models were fine-tuned during 3 epochs using a linear scheduler with a 5% warmup ratio and FP16 precision. The specific hyperparameters for each model size are detailed in Table 3. Figure 4 demonstrates the fine-tuning process through the loss curves, showing no signs of overfit.

TABLE 3. Whisper model fine-tuning hyperparameters.

Model Size	Number of Parameters	Learning Rate	Batch Size	Accumulation Steps
Small	244 M	1e-5	4	8
Medium	769 M	1e-6	8	4
Large v3	1550 M	1e-6	8	4

To maintain an effective batch size of 32 across all Whisper models, we carefully balanced the trade-offs between batch size and accumulation steps, given the constraints of the available GPU memory (96GB on an AWS g5.12xlarge instance). For the medium and large Whisper models,

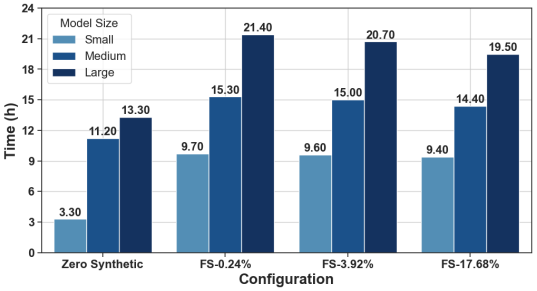

FIGURE 5. Runtime in hours for conducting each experiment, based on the specified model size.

gradient checkpointing allowed us to increase the per-GPU batch size while managing memory usage. Additionally, for the largest Whisper model, we employed DeepSpeed ZeRO stage 2 to further optimize memory consumption, enabling efficient training with the desired effective batch size. The runtime for each configuration is shown in Figure 5. In total, it was required 166,1 training hours across different model architectures to conduct this research

V. RESULTS

We evaluated the performance of the synthetic data filtering methodology on two Portuguese datasets: PSFB and FLEURS.

On the PSFB dataset, using the small Whisper model, a moderate level of synthetic data filtering (FS−3.92%) yielded the best performance, with a WER reduction of 1.24 percentage points compared to the model trained on real data only. Remarkably, using no synthetic data led to better results than overly aggressive filtering (FS−17.68%), reducing WER by 1.17 percentage points compared to the latter. This highlights the small model's sensitivity to the removal of potentially valuable training examples, suggesting that a balance between data quantity and quality, achieved through moderate filtering, is beneficial for smaller models, potentially mitigating the impact of noisy or less representative synthetic samples

In contrast, the large and medium models exhibited a different pattern, achieving their lowest WER when only a small amount of synthetic data was filtered out (FS−0.24%). Notably, the absence of synthetic data resulted in the worst performance for both models, underscoring the importance of data augmentation even for larger architectures. This suggests that these models benefit from the increased diversity provided by synthetic data, even with minimal filtering, potentially due to their enhanced capacity to discern relevant patterns and discard noisy or less relevant samples.

On the FLEURS dataset, on the large Whisper model, a rigorous filtering approach (FS−17.68%) yielded the best performance, reducing WER by 0.45 percentage points compared to minimal filtering (FS−0.24%). Interestingly, the model trained only on real data outperformed the minimally filtered model, underscoring the large model's sensitivity to the quality of synthetic data. This suggests that large models, with their increased capacity to capture intricate patterns, benefit most from a highly curated dataset where lower-quality synthetic samples are removed.

For the small and medium Whisper models, a moderate filtering level (FS−3.92%) proved optimal, reducing WER by 0.80 and 0.30 percentage points, respectively, compared to using no synthetic data. In contrast to the large model, more aggressive filtering did not yield further improvements for these smaller models. This indicates that smaller models might benefit from the increased diversity offered by a larger amount of synthetic data, even if it includes some less relevant samples.

TABLE 4. WER results for FLEURS and PSFB datasets across different whisper models and filtering configurations without text normalization.

Dataset	Model	FS-0.24%	FS-3.92%	FS-17.68%	Zero Synthetic
PSFB	Small	35.60	**34.05**	36.46	35.29
	Medium	**32.33**	33.11	33.10	33.90
	Large V3	**28.19**	28.72	28.29	29.26
FLEURS	Small	18.90	**18.53**	18.57	19.33
	Medium	14.54	**14.44**	14.46	14.74
	Large V3	11.28	11.09	**10.83**	10.92

Table 5 compares the performance of the fine-tuned Whisper models against their pre-trained counterparts on the FLEURS dataset, evaluating the WER on both normalized (lowercase, punctuation removed) and non-normalized text. As expected, all fine-tuned models demonstrate a substantially reduced normalized text WER compared to their pre-trained versions across all model sizes.

TABLE 5. WER results for FLEURS for the fine-tuned model versus pretrained model with and without text normalization.

Model Size	Model Type	WER	
		Normalized	Non-Normalized
Small	Pretrained[2]	10.87	**15.43**
	FS-17.68%	10.45	18.57
	FS-3.92%	**10.34**	18.53
	FS-0.24%	10.58	18.90
	Zero Synthetic	10.90	19.32
Medium	Pretrained[3]	8.62	**12.65**
	FS-17.68%	6.58	14.46
	FS-3.92%	**6.57**	14.44
	FS-0.24%	6.58	14.54
	Zero Synthetic	6.97	14.74
Large V3	Pretrained[4]	7.70	11.78
	FS-17.68%	4.73	**10.83**
	FS-3.92%	**4.65**	11.09
	FS-0.24%	4.80	11.28
	Zero Synthetic	4.86	10.92

Small and medium models demonstrated optimal performance with moderate filtering (FS−3.92%) for normalized

[2]https://huggingface.co/openai/whisper-small
[3]https://huggingface.co/openai/whisper-medium
[4]https://huggingface.co/openai/whisper-large-v3

text. However, for non-normalized text, the pre-trained models consistently outperformed their fine-tuned versions, regardless of the use of synthetic data. This suggests that fine-tuning smaller models with either real or synthetic data may be less beneficial when targeting strong performance on non-normalized text. These models, with their limited capacity, likely struggle to effectively utilize the additional information from diverse training data, particularly when inconsistencies such as punctuation errors or unrefined samples are present. Nevertheless, both models show improvements when fine-tuned with a mix of synthetic and real audio data, especially with more aggressive filtering, which helps mitigate the impact of lower-quality data.

In contrast, the large model showed a preference for aggressive filtering (FS−17.68%) on non-normalized text, demonstrating its ability to discern and benefit from high-quality synthetic data while being more sensitive to lower-quality samples. This can be attributed to its larger architecture and higher number of parameters, allowing it to capture useful patterns even in noisy data, such as incorrect punctuation or capitalization. For normalized text, the Large V3 model performed best with moderate filtering (FS−3.92%). However, on non-normalized text, the model trained exclusively on real data (Zero Synthetic) outperformed those trained with minimally filtered (FS−0.24%) and moderately filtered (FS−3.92%) synthetic data, highlighting the importance of rigorous filtering for large models in this scenario.

This performance drop is likely due to limitations in our current filtering model, particularly for FS−0.24% and FS−3.92%, which fail to remove certain erroneous samples, such as:

- **Sample 1:**"A lesão espinhalcompressiva foi realizada com insuflação de quinze l, de solução salina, do cuff do cateterde forgat n. dois fr.no espação epidural t oito, durante cinco minutos.os animais foram distribuídos aleatoriamente…". Corresponding in English:"The spinalcompression injury was carried out by blowing fifteen l, of saline solution, through the cuff of the forgat n. two fr. catheter into the epidural space t eight for five minutes.s animals were randomly allocated…"
- **Sample 2:**"Em morus nigra quantificou-se dois mil trezentos e vinte e três,noventa cento e quarenta e cinco,trinta e cinco g.g um e mil quatrocentos e quarenta e seis,trinta e seis cinquenta e nove,zero g.g um, de quercetina e canferol, respectivamente…". Corresponding in English: "In morus nigra were quantified two thousand three hundred and twenty-three, ninety one hundred and forty-five, thirty-five g.g. one and one thousand four hundred and forty-six, thirty-six fifty-nine, zero g.g. one quercetin and kaempferol respectively…"
- **Sample 3:** "A densidade média para a população de uca sp. variou de dezoito,um um,vinte e nove a sessenta e dois,vinte e dois um,cinquenta e um ind.m, para uca thayeri de três,sessenta e sete zero,sessenta e três…". Corresponding in English: "The average density for the uca sp. population varied from eighteen, one, twenty-nine to sixty-two, twenty-two, one, fifty-one ind.m, for uca thayeri from three, sixty-seven, zero, sixty-three…"

These samples show limited variation and numerous inconsistencies, such as punctuation and formatting errors, which introduce noise and reduce the overall quality of the training data. To address this issue, future work will focus on improving the filtering model by expanding the training dataset and incorporating a robust preprocessing step for the text data used in generating synthetic samples.

VI. CONCLUSION

In this work, we presented a novel methodology for enhancing Automatic Speech Recognition (ASR) performance by utilizing contrastive learning to filter synthetic audio data. The proposed approach addresses the challenge of incorporating synthetic data effectively, particularly in scenarios where the target domain exhibits unique linguistic characteristics.

Experiments conducted on the FLEURS and PSFB datasets confirm the effectiveness of the methodology. We observed consistent improvements in WER across different Whisper model sizes when using filtered synthetic data. In particular, the results highlight the importance of adapting the synthetic data filtering strategy to both the specific model architecture and the characteristics of the target domain.

However, we observed limitations, particularly with the smaller models in the non-normalized setup, likely due to the inaccuracies in the synthetic data used for fine-tuning. This highlights one of the risks of using synthetic data, as discussed in [102]. Smaller models may internalize these erroneous patterns, leading to biased predictions and negatively impacting the model's overall performance and reliability. Moreover, extending fine-tuning experiments to new settings or environments, especially with larger models or more extensive synthetic datasets, could substantially increase computational costs.

While the results are encouraging, there are several directions for future work. One idea is to expand the diversity of data used to train the cross-modal embedding model, which could lead to more robust audio-text representations. Another promising direction is to explore how different model architectures perform when trained only on synthetic data, enabling a comparative analysis that could reveal specific strengths or weaknesses across models. Additionally, it is important to explore how the gender of the speaker in the synthetic audio influences the models' performance and evaluate the quality of the source datasets.

REFERENCES

[1] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, "Making more of little data: Improving low-resource automatic speech recognition using data augmentation," 2023, *arXiv:2305.10951*.

[2] N. Rossenbach, A. Zeyer, R. Schlüter, and H. Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7069–7073.
[3] C. Du and K. Yu, "Speaker augmentation for low resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7719–7723.
[4] S. Liu, L. Sari, C. Wu, G. Keren, Y. Shangguan, J. Mahadeokar, and O. Kalinli, "Towards selection of text-to-speech data to augment ASR training," 2023, *arXiv:2306.00998*.
[5] J. Huang, Y. Bai, Y. Cai, and W. Bian, "A study on the adverse impact of synthetic speech on speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 33, Apr. 2024, pp. 10266–10270.
[6] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.
[7] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2019, *arXiv:1912.06670*.
[8] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," 2020, *arXiv:2012.03411*.
[9] E. E. David, "Ears for computers," *Sci. Amer.*, vol. 192, no. 2, pp. 92–98, Feb. 1955.
[10] B. Lowerre and R. Reddy, "The harpy speech recognition system: Performance with large vocabularies," *J. Acoust. Soc. Amer.*, vol. 60, pp. 10–11, Nov. 1976.
[11] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 1, pp. 67–72, Feb. 1975.
[12] D. Yu and L. Deng, *Automatic Speech Recognition*, vol. 1. Berlin, Germany: Springer, 2016.
[13] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.
[15] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. ASSPM-3, no. 1, pp. 4–16, Jan. 1986.
[16] D. A. Reynolds, "Gaussian mixture models," *Encyclopedia Biometrics*, vol. 741, pp. 659–663, Jul. 2009.
[17] B.-H. Juang, S. Levinson, and M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 2, pp. 307–309, Mar. 1986.
[18] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 11, Aug. 1986, pp. 49–52.
[19] B. Mor, S. Garhwal, and A. Kumar, "A systematic review of hidden Markov models and their applications," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 1429–1448, May 2021.
[20] W. Mcculloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biol.*, vol. 52, nos. 1–2, pp. 99–115, 1990.
[21] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2007, pp. 4–757.
[22] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, Jun. 2000, pp. 1635–1638.
[23] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP)*, Oct. 2000, pp. 1–8.
[24] L. Tóth, "A hierarchical, context-dependent neural network architecture for improved phone recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5040–5043.
[25] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980.
[26] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proc. Interspeech*, 2015, pp. 11–15, doi: 10.21437/Interspeech.2015-3.
[27] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4277–4280.
[28] G. D. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.
[29] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.
[30] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.
[31] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 315–320.
[32] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci. USA*, vol. 79, no. 8, pp. 2554–2558, 1982.
[33] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[35] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[37] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey," *Inf. Fusion*, vol. 109, Sep. 2024, Art. no. 102422.
[38] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.
[39] H. Yang, M. Zhang, S. Tao, M. Ma, and Y. Qin, "Chinese ASR and NER improvement based on whisper fine-tuning," in *Proc. 25th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2023, pp. 213–217.
[40] C. Sicard, K. Pyszkowski, and V. Gillioz, "Spaiche: Extending state-of-the-art ASR models to Swiss German dialects," 2023, *arXiv:2304.11075*.
[41] P. Xie, X. Liu, Z. Chen, K. Chen, and Y. Wang, "MADGF: Multi-agent data generation framework," 2023, *arXiv:2310.17953*.
[42] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, and J. Hoffman, "SeamlessM4T: Massively multilingual & multimodal machine translation," 2023, *arXiv:2308.11596*.
[43] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder–decoder pre-training for spoken language processing," 2021, *arXiv:2110.07205*.
[44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
[45] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource ASR," *EURASIP J. Audio, Speech, Music Process.*, vol. 2024, no. 1, p. 29, Jun. 2024.
[46] N. Rossenbach, B. Hilmes, and R. Schlüter, "On the effect of purely synthetic training data for different automatic speech recognition architectures," 2024, *arXiv:2407.17997*.
[47] J. C. Vásquez-Correa, H. Arzelus, J. M. Martin-Doñas, J. Arellano, A. Gonzalez-Docasal, and A. Álvarez, "When whisper meets tts: Domain adaptation using only synthetic speech data," in *Proc. Int. Conf. Text, Speech, Dialogue*. Cham, Switzerland: Springer, Aug. 2023, pp. 226–238.
[48] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Data augmentation for ASR using TTS via a discrete representation," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 68–75.
[49] Z. S. Harris, "Distributional structure," *Word*, vol. 10, nos. 2–3, pp. 146–162, 1954.
[50] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[51] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, A. Moschitti, B. Pang, and W. Daelemans, Eds., 2014, pp. 1532–1543.

[52] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016, *arXiv:1409.0473*.
[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," 2023, *arXiv:1706.03762*.
[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019, *arXiv:1810.04805*.
[55] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[56] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "AlBERT: A lite BERT for self-supervised learning of language representations," 2020, *arXiv:1909.11942*.
[57] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2021, *arXiv:2006.03654*.
[58] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.
[59] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate features and Adaboost for music classification," *Mach. Learn.*, vol. 65, pp. 473–484, Dec. 2006.
[60] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 258–261, Feb. 2004.
[61] S.-A. Selouani, M. Sidi Yakoub, and D. O'Shaughnessy, "Alternative speech communication system for persons with severe speech disorders," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 1–12, Dec. 2009.
[62] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1692–1707, Sep. 2010.
[63] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019, *arXiv:1807.03748*.
[64] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.
[65] W.-N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.
[66] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
[67] K. Q. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 18, Y. Weiss, B. Schölkopf, and J. Platt, Eds., Cambridge, MA, USA: MIT Press, 2005, pp. 1–8.
[68] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Curran Associates: Red Hook, NY, USA, 2016, pp. 1857–1865.
[69] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[70] S. Tonekaboni, D. Eytan, and A. Goldenberg, "Unsupervised representation learning for time series with temporal neighborhood coding," 2021, *arXiv:2106.00750*.
[71] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," 2019, *arXiv:1808.06670*.
[72] H. Al-Tahan and Y. Mohsenzadeh, "CLAR: Contrastive learning of auditory representations," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2530–2538.
[73] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Jun. 2020, pp. 3875–3879.
[74] M. T. Islam and S. Nirjon, "SoundSemantics: Exploiting semantic knowledge in text for embedded acoustic event classification," in *Proc. 18th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2019, pp. 217–228.
[75] L. Wang and A. van den Oord, "Multi-format contrastive learning of audio representations," 2021, *arXiv:2103.06508*.
[76] Y. Mori, H. Takahashi, and R. I. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proc. 1st Int. Workshop Multimedia Intell. Storage Retr. Manage.*, 1999, pp. 1–9.
[77] A. Quattoni, M. Collins, and T. Darrell, "Learning visual representations using images with captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[78] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
[79] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 740–755.
[80] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.
[81] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *Int. J. Comput. Vis.*, vol. 123, pp. 94–120, May 2016.
[82] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.
[83] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.
[84] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.
[85] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, 2015, pp. 1015–1018.
[86] A. Bapna, Y. A. Chung, N. Wu, A. Gulati, Y. Jia, J. H. Clark, M. Johnson, J. Riesa, A. Conneau, and Y. Zhang, "SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training," 2021, *arXiv:2110.10329*.
[87] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[88] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 976–980.
[89] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning robust audio representations from clip," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4563–4567.
[90] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[91] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[92] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 24206–24221.
[93] L. Wang, P. Luc, A. Recasens, J.-B. Alayrac, and A. van den Oord, "Multimodal self-supervised learning of general audio representations," 2021, *arXiv:2104.12807*.
[94] P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, and C. Zhou, "One-peace: Exploring one general representation model toward unlimited modalities," 2023, *arXiv:2305.11172*.
[95] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022, *arXiv:2212.04356*.
[96] A. J. Herdeiro and P. Barbosa, "O fenómeno do queísmo no falar bracarense: Um estudo sociolinguístico," Diacrítica, Série Ciências da Linguagem, Universidade do Minho, Centro de Estudos Humanísticos (CEHUM), 2015.
[97] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," 2022, *arXiv:2208.12208*.
[98] R. Santos, J. Rodrigues, L. Gomes, J. Silva, A. Branco, H. L. Cardoso, T. F. Osório, and B. Leite, "Fostering the ecosystem of open neural encoders for Portuguese with Albertina PT* family," 2024, *arXiv:2403.01897*.

[99] L. Barrault et al., "Seamless: Multilingual expressive and streaming speech translation," 2023, *arXiv:2312.05187*.
[100] F. Soares, G. H. Yamashita, and M. J. Anzanello, "A parallel corpus of theses and dissertations abstracts," in *Proc. 13th Int. Conf. Comput. Process. Portuguese Lang. (PROPOR)* (Lecture Notes in Computer Science), vol. 11122, Canela, Brazil. Cham, Switzerland: Springer, Sep. 2018, pp. 345–352.
[101] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," 2022, *arXiv:2205.12446*.
[102] S. Hao, W. Han, T. Jiang, Y. Li, H. Wu, C. Zhong, Z. Zhou, and H. Tang, "Synthetic data in AI: Challenges, applications, and ethical implications," 2024, *arXiv:2401.01629*.

YURIY PEREZHOHIN is currently pursuing the Ph.D. degree in data science with NOVA IMS, Lisbon, Portugal, where he is developing new methodologies for reducing the computational resources of large language models in their training or inference process. He has developed during the master's thesis an innovative approach to tackle text-to-SQL tasks, where the model is free of training and adaptable to new domains. He also lectures practical classes on deep learning at NOVA IMS. His research interests include natural language processing, code generation, and cross-modal embeddings.

TIAGO SANTOS is currently pursuing the M.Sc. degree in data science with NOVA IMS, Lisbon, Portugal. His master's thesis focuses on large language models for semantic table interpretation, where he developed a state-of-the-art model that outperformed existing approaches. His research interests include natural language processing (NLP) and automatic speech recognition (ASR).

VICTOR COSTA received the master's degree in data science. He is currently pursuing the Ph.D. degree with NOVA IMS. He has developed systems to identify at-risk Portuguese families and extract legal document information using deep learning. He has taught courses in computational thinking and optimization at NOVA IMS and has extensive experience in software development, working as a User Experience Designer, a Software Engineer, and the Product Manager for various tech companies. He is also a Machine Learning Engineer. His research interests include generating labeled data at scale for training ML models.

FERNANDO PERES is currently pursuing the Ph.D. degree. He is also the CTO and AI Engineer at MyNorth Group and an Invited Professor at NOVA IMS, Lisbon, Portugal. His Ph.D. thesis focuses on integrating context and semantic knowledge for AI and metaheuristics. His research interests include natural language processing (NLP), generative AI, combinatorial optimization, agents, and explainable AI.

MAURO CASTELLI received the Ph.D. degree in computer science from the University of Milano-Bicocca, Milan, Italy, in 2012. He is currently an Associate Professor of computational intelligence and machine learning with the NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, Lisbon, Portugal. He is also the Director of the bachelor's program in data science at NOVA IMS. He has published his research in a variety of top-quality academic outlets, such as IEEE TRANSACTIONS ON CYBERNETICS and IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. He has published more than 200 contributions in international conferences and journals in the field of machine learning. He led several national and international research projects on artificial intelligence and supervised more than 130 master's students and nine Ph.D. candidates.