BRSpeech-DF: A Deep Fake Synthetic Speech Dataset for Portuguese Zero-Shot TTS

Alexandre C. Ferro Filho[1, 2*], Rafaello Virgilli[1, 2], Lucas A. Souza[1, 2], Frederico S. de Oliveira[1, 3], Marcelo Henrique L. Ferreira[2], Daniel Tunnermann[1, 2], Gustavo R. Oliveira[2], Anderson S. Soares[1, 2], Arlindo R. Galvão Filho[1, 2]

[1] Advanced Knowledge Center in Immersive Technologies (AKCIT)
[2] Federal University of Goiás (UFG)
[3] Federal University of Mato Grosso (UFMT)

Abstract

The detection of audio deepfakes (ADD) has become increasingly important due to the rapid evolution of generative speech models. However, progress in this field remains uneven across languages, particularly for low-resource languages like Portuguese, which lack high-quality datasets. In this paper, we introduce **BRSpeech-DF**, the first publicly available ADD dataset for Portuguese, encompassing both Brazilian and European variants. The dataset contains over 458,000 utterances, including a smaller portion of real speech from 62 speakers and a large collection of synthetic samples generated using multiple zero-shot text-to-speech (TTS) models, each conditioned on the original speaker's voice. By providing this resource, our objective is to support the development of robust, multilingual detection systems, thereby advancing equity in speech forensics and security research. BRSpeech-DF addresses a significant gap in annotated data for underrepresented languages, facilitating more inclusive and generalizable advancements in synthetic speech detection.

Our dataset and codes are publicly available[1].

1 Introduction

The increasing sophistication of generative models has resulted in a proliferation of highly realistic synthetic media (Xie et al., 2025), including speech deepfakes. While such technologies offer promising opportunities for accessibility and human-computer interaction, they also pose significant threats to societal trust, privacy, and security (Kharvi, 2024). The potential for audio deepfakes to be weaponized for impersonation, fraud, and misinformation is particularly concerning in scenarios where voice-based authentication or content verification plays a critical role.

Recent breakthroughs in zero-shot TTS have further lowered the barriers to generating speech in arbitrary voices, requiring only a few seconds of reference audio to convincingly imitate a speaker without any fine-tuning (Li et al., 2024b; Bang and Chun, 2023). These advancements make it increasingly easy to generate realistic audio for malicious purposes, raising the urgency for robust audio deepfake detection (ADD) systems.

In parallel, the research community has made notable strides in developing detection techniques for deepfakes, employing both features and deep learning methods to distinguish synthetic from real audio (Delgado et al., 2021; Wang et al., 2025). However, the majority of existing datasets and benchmarks focus on high-resource languages, primarily English, creating a linguistic bias in the training and evaluation of detection systems (Li et al., 2024a).

Moreover, the combination of zero-shot synthesis capabilities and the lack of annotated deepfake corpora in underrepresented languages presents a unique and actual research challenge. Without appropriate datasets, it is difficult to assess the effectiveness of detection models or develop defenses that are culturally and linguistically inclusive.

To address this gap, we introduce **BRSpeech-DF**, the first publicly available dataset designed for audio deepfake detection in Portuguese. By leveraging modern zero-shot TTS models to generate high-quality synthetic utterances, paired with corresponding real speech samples, our dataset enables the development and evaluation of detection systems in a linguistically diverse setting. We hope this resource will encourage further research on ADD for low-resource languages and foster the creation of more equitable and robust speech security technologies.

* Corresponding author: alexandre_ferro@discente.ufg.br

[1] https://github.com/AKCIT-Speech/BRSpeech-DF-Dataset

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35122–35127
November 4-9, 2025

2 Dataset

The BRSpeech-DF dataset comprises a total of 458,411 audio files in Brazilian and European Portuguese, encompassing 160 hours of natural speech and 823 hours of synthetic speech, distributed between 62 real speakers (37 men and 25 women) whose voices were cloned using zero-shot models. For each sentence in the real speech set, synthetic versions were generated using all 5 zero-shot models, with the reference speaker's own voice always maintained. BRSpeech-DF is organized into three mutually exclusive splits - training, validation and test - defined by speaker separation, in order to avoid voice leakage between the sets. Table 1 shows the number of real and synthetic samples in each split.

Split	Bona fide	Spoof	N. Speakers
Train	73,949	369,625	42
Validation	1,158	5,788	10
Test	1,316	6,575	10
Total	76,423	381,988	62

Table 1: Distribution of real and synthetic samples and number of speaker by split.

2.1 Real Speech Sources

BRSpeech is a dataset comprising audiobooks sourced from the public domain books of Project Gutenberg, which are then read by volunteers from the LibriVox[2] project. The dataset comprises recordings in Brazilian and European Portuguese. Each recording has undergone a speech enhancement procedure to remove acoustic artifacts, including light noise, hissing, echoes, and light reverberation. This process was executed utilizing the Vocos model[3], applied as a denoiser. The model functions at a 48 kHz sampling rate and was previously trained exclusively for this task on a proprietary dataset.

Part of the bona fide data in BRSpeech-DF overlaps with the CML-TTS corpus, since both were sourced from LibriVox audiobooks. We found 35,173 identical utterances, covering 50 speakers. This represents 46% of the BRSpeech-DF bona fide set and nearly all of CML-TTS. Despite this overlap, BRSpeech-DF expands the scale to 62 speakers and standardizes recordings at 48 kHz, while CML-TTS is released at 24 kHz.

[2] https://librivox.org/

[3] https://github.com/gemelo-ai/vocos

2.2 Construction of the Synthetic Data

Zero-Shot Models Employed. To generate the synthetic versions of each sentence, five zero-shot speech synthesis models were employed: **Fish Speech** (Liao et al., 2024), **XTTS** (Casanova et al., 2024), **F5-TTS** (Chen et al., 2024), **YourTTS** (Casanova et al., 2022) and **ToucanTTS** (Lux et al., 2024).

Each model was configured with its official release, and the original implementations were kept without structural modifications, ensuring comparability with literature results.

Synthesis Pipeline. The generation of synthetic samples was based directly on the sentences in the original speech corpus. Initially, all recordings lacking an associated textual transcription were discarded, ensuring semantic integrity between audio and text during synthesis. For each remaining real sentence, and for each of the five zero-shot models utilized, cloning was carried out using the original audio itself as the voice reference and the corresponding text as the textual input.

This procedure ensured that the speaker's vocal identity was preserved in all synthetic versions of the same sentence, which is essential for creating convincing examples of deepfakes.

To ensure consistency in the evaluation pipeline, all audio was standardized to **24 kHz**, including resampled outputs from FishSpeech (48 kHz) and YourTTS (16 kHz).

During the synthesis pipeline, some utterances were discarded to mantein consistency and avoid problematic cases. Very short samples were excluded from XTTS due to generation errors related to duration constraints. In addition, all utterances longer than 35 seconds were removed from the final dataset to mitigate evident outliers in duration, particularly observed in Fish Speech outputs. This filtering step made sure that the resulting corpus maintains a more balanced and realistic distribution of utterance lengths.

2.3 Dataset Analysis

While uTMOS was originally trained on English data, which may affect its applicability to Portuguese, we employed it as a practical alternative to enable consistent comparisons across TTS models. No MOS predictors trained specifically for Portuguese are currently available. Moreover, recent studies (Sellam et al., 2023) suggest that paralinguistic features allow some MOS predictors to

Model	uTMOS ↑	Spk. Sim. ↑	SI_SDR (dB) ↑	WER (%) ↓
F5-TTS	3.39	0.71	20.86	14.13
XTTS	3.37	0.71	25.97	5.5
ToucanTTS	3.20	0.60	26.50	19.18
Fish-Speech	3.11	0.76	22.33	11.43
YourTTS	2.58	0.41	16.92	14.21
Bona fide	3.44	-	-	9.11

Table 2: Mean results of perfomance metrics for the selected models

generalize reasonably well across languages.

The evaluation of synthetic audio quality reveals significant variations among naturalness, speaker similarity, noise levels, and intelligibility.

The uTMOS scores indicate that F5-TTS (3.395) and XTTS (3.377) produce the most natural-sounding speech. In contrast, YourTTS (2.589) lags significantly, which aligns with observations of its noisier output. Fish Speech achieves the highest speaker similarity score (0.764), surpassing models with high naturalness scores, such as F5-TTS (0.713). However, ToucanTTS (0.603) and YourTTS (0.418) demonstrate significantly lower similarity between synthesized speech and the original speaker's voice. The evaluation of synthetic audio quality reveals significant variations in naturalness, speaker similarity, noise levels, and intelligibility.

In a further analysis of signal quality, YourTTS's lower SI-SDR (16.93 dB) indicates the presence of background noise or spectral distortions. In contrast, ToucanTTS achieves the highest SI-SDR (26.51 dB), reflecting cleaner signal reconstruction. XTTS follows closely with 25.98 dB, suggesting minor artifacts that do not significantly degrade overall quality.

However, intelligibility, measured by WER , further exposes weaknesses. ToucanTTS (19.18%), attributed to unstable prosody and barely comprehensible Portuguese words, and YourTTS (14.21%) exhibit comparatively poorer performance. In contrast, XTTS (5.5%) and Fish Speech (11.43%) demonstrate superior intelligibility.

Models like XTTS, F5-TTS, and Fish Speech strike a balance between naturalness and intelligibility. However, ToucanTTS notably underperforms in terms of intelligibility, despite its good signal quality. YourTTS performs poorly in both areas.

As shown below, despite the variability in audio quality , none of the synthetic audios were reliably detectable by the spoofing detector. Even lower-quality synthetic speech (e.g., YourTTS) retains sufficient acoustic coherence to bypass current detection systems, raising concerns about the need to improve adversarial robustness in voice authentication pipelines.

3 Benchmarking and Experiments

In this section, we evaluate how well existing deepfake detection models perform on our new dataset, **BRSpeech-DF**. We aim to demonstrate two main points: (i) to what extent standard detectors trained on other languages generalize to Brazilian Portuguese without retraining, and (ii) why a dedicated Portuguese dataset is valuable for future research. To fairly evaluate these models, we calibrated each detector's decision threshold using their respective original evaluation datasets.

3.1 Experimental Setup

For our experiments, we report all results using the *test* split (see Table 1), while reserving the *dev* split for future experiments such as fine-tuning. To evaluate performance, we report accuracy (ACC) at each model's original calibration threshold and equal error rate (EER) obtained by sweeping the threshold on the BRSpeech-DF test scores. We selected four well-known detection models:

- AASIST (weon Jung et al., 2021), calibrated on ASVspoof 2019 LA (Wang et al., 2020) (threshold = 1.49);
- SLS-ECAPA (Zhang et al., 2024), calibrated on ASVspoof 2021 DF (Yamagishi et al., 2021) (threshold = -11.52);
- SSL Anti-spoof (Tak et al., 2022), calibrated on ASVspoof 2021 DF (Yamagishi et al., 2021) (threshold = -3.53);
- XLSR-MAMBA (Xiao and Das, 2024), calibrated on ASVspoof 2021 DF (threshold = -4.13).

3.2 Overall Results

Table 3 summarizes the main results.The SLS-ECAPA model performed the best (EER = 30.67%), followed by SSL Anti-spoof (EER = 35.38%). AASIST (EER = 48.12%) and XLSR-MAMBA (EER = 63.03%) performed considerably worse, with XLSR-MAMBA's performance close to random guessing. These results confirm a significant mismatch when transferring detection models trained primarily on English data to Brazilian Portuguese.

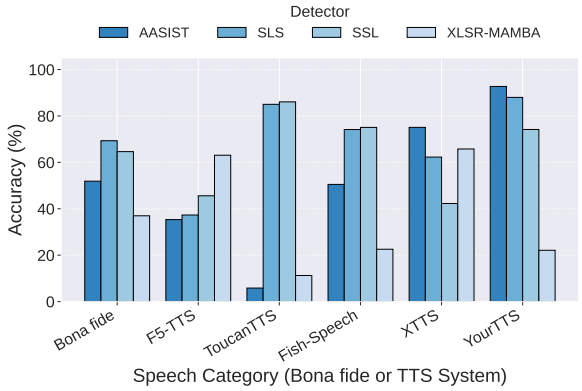


Figure 1: Accuracy per speech category. Detection rate = TNR for bonafide, TPR for spoof. Categories are sorted by average detection difficulty (lower average accuracy indicates harder category).

Table 3 also lists the detectors' reference performance on their original benchmarks (ASVspoof 2019 LA for AASIST and ASVspoof 2021 DF for the others). It is evident that all models experience a substantial performance drop on the BRSpeech-DF dataset compared to their original evaluation sets. This contrast highlights the challenge of generalizing deepfake detection models across languages and domains.

Detector	ACC (%)	EER (%)	Ref. EER (%)
AASIST	63.03	48.12	0.83
SLS-ECAPA	75.97	30.67	1.92
SSL Anti-spoof	70.22	35.38	2.85
XLSR-MAMBA	16.72	63.03	1.88

Table 3: Overall performance on BRSpeech-DF test set. ACC is measured at the original calibration threshold.

3.3 Performance by TTS system

To understand how models perform against different synthesis techniques, we analyzed their detection rates (DR) for bona fide samples and for spoofed samples generated by each of the five TTS systems, using the individual EER thresholds for each model. These results are detailed in Figure 1. Detection rate corresponds to TNR for bona fide samples and TPR for spoofed samples.

Overall, SLS and SSL demonstrated a more robust performance across various TTS systems compared to AASIST and XLSR-MAMBA, highlighting that different TTS architectures produce artifacts that current detectors exploit to varying degrees, underscoring the need for diverse training data and robust models for generalized deepfake detection.

3.4 Score Distributions and Threshold Analysis

Figure 2 compares the distribution of model scores for bona fide and spoofed samples. For the AASIST model (Figure 2a), the original threshold falls within a large overlap between real and fake samples, explaining its high error rate. The SLS-ECAPA model (Figure 2b), in contrast, demonstrates clearer separation between classes, which aligns with its better overall performance. The SSL model (Figure 2c) also shows reasonable separation.

3.5 Key Observations

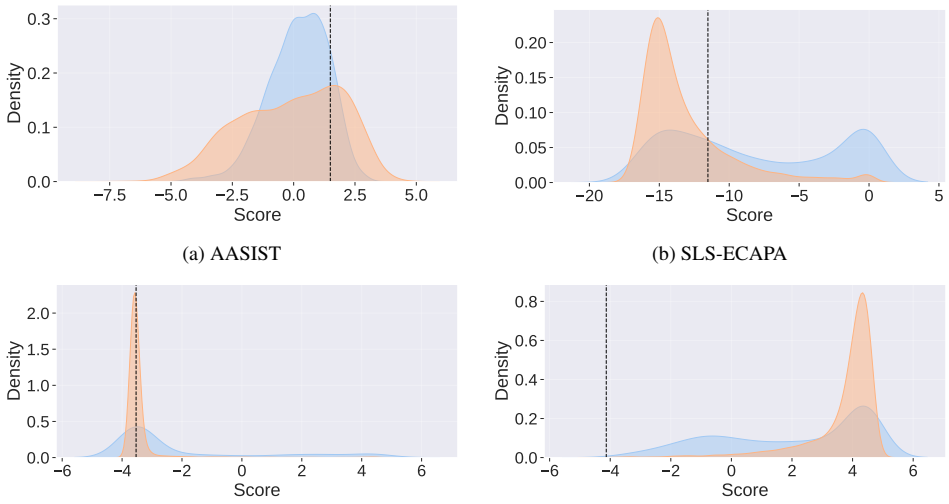
Our results reveal that deepfake detectors trained on English datasets generalize poorly to Brazilian Portuguese. Threshold tuning alone proved ineffective (e.g., AASIST), and certain TTS systems, such as `F5-TTS`, produced speech that was particularly hard to detect. These insights reinforce the need for language-specific resources like **BRSpeech-DF** to enable robust multilingual detection.

4 Conclusion

In conclusion, our work introduces **BRSpeech-DF**, a comprehensive Portuguese audio deepfake dataset comprising over 458,000 samples (983 hours) of real and zero-shot TTS-generated speech. Our critical findings reveal that state-of-the-art detectors trained on English data perform poorly on Portuguese data (EER: 30.67%-63.03%), and even low-quality synthetic samples evade detection. This exposes systemic vulnerabilities in cross-lingual generalization. These results underscore the urgent need for language-specific resources to develop robust detection frameworks. By providing an open, linguistically diverse dataset, we enable equitable research into multilingual defenses against evolving synthetic threats and bridge gaps in security for underrepresented languages.

5 Limitations

This work presents some limitations that should be addressed in future research. All detectors were evaluated using their original decision thresholds, without any calibration for Portuguese speech, which likely impacted the final accuracy and may have influenced the drop in performance. The evaluation was restricted to synthetic speech generated



(a) AASIST (b) SLS-ECAPA

(c) SSL Anti-spoof (d) XLSR-MAMBA

Figure 2: Score distributions for bona fide (light blue) and spoofed (coral) samples. Vertical dashed lines indicate the original/default calibration thresholds: AASIST (1.49), SLS-ECAPA (-11.52), SSL Anti-spoof (-3.53), and XLSR-MAMBA (-4.13).

by only five zero-shot TTS models, and all audio samples were clean, without compression or environmental noise, limiting the realism of the test conditions. Furthermore, the XLSR-MAMBA model exhibited notably poor performance, and the current experimental setup does not allow a clear explanation for this result. Additional tests are necessary to better understand its behavior and potential limitations in cross-lingual scenarios.

In addition, part of BRSpeech-DF overlaps with CML-TTS, sharing approximately 35k utterances and 50 speakers. Although our dataset includes additional speakers and higher-resolution audio, this overlap should be considered when designing cross-corpus evaluations.

Another limitation concerns speaker diversity: since all bona fide data comes from LibriVox audiobooks, the speaking style is limited to read speech. The absence of conversational or spontaneous speech may affect the generalization of detection models trained solely on our corpus.

Acknowledgments

This work has been fully funded by the project Research and Development of Algorithms for Construction of Digital Human Technological Components supported by the Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

References

Chae-Woon Bang and Chanjun Chun. 2023. Effective zero-shot multi-speaker text-to-speech technique using information perturbation and a speaker encoder. *Sensors (Basel)*, 23(23).

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pages 4978–4982.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, and Junichi Yamagishi. 2021. Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *Preprint*, arXiv:2109.00535.

Prakash L. Kharvi. 2024. Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media . *IEEE Security & Privacy*, 22(04):115–122.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2024a. Audio anti-spoofing detection: A survey. *Preprint*, arXiv:2404.13914.

Yinghao Aaron Li, Xilin Jiang, Cong Han, and Nima Mesgarani. 2024b. Styletts-zs: Efficient high-quality zero-shot text-to-speech synthesis with distilled time-varying style diffusion. *Preprint*, arXiv:2409.10058.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *Preprint*, arXiv:2411.01156.

Florian Lux, Sarina Meyer, Lyonel Behringer, Frank Zalkow, Phat Do, Matt Coler, Emanuël A. P. Habets, and Ngoc Thang Vu. 2024. Meta Learning Text-to-Speech Synthesis in over 7000 Languages. In *Interspeech*. ISCA.

Thibault Sellam, Ankur Bapna, Joshua Camp, Diana Mackinnon, Ankur P. Parikh, and Jason Riesa. 2023. Squid: Measuring speech naturalness in many languages. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop*.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, and 10 others. 2025. Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Preprint*, arXiv:2502.08857.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, and 21 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Preprint*, arXiv:1911.01601.

Jee weon Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2021. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. *Preprint*, arXiv:2110.01200.

Yang Xiao and Rohan Kumar Das. 2024. XLSR-Mamba: A dual-column bidirectional state space model for spoofing attack detection. *arXiv preprint arXiv:2411.10027*.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025. Towards controllable speech synthesis in the era of large language models: A survey. *Preprint*, arXiv:2412.06602.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *Preprint*, arXiv:2109.00537.

Qishan Zhang, Shuangbing Wen, and Tao Hu. 2024. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 6765–6773, New York, NY, USA. Association for Computing Machinery.

A Appendix: Open Source Tools

Zero-shot TTS models:

- Fish Speech; https://huggingface.co/fishaudio/fish-speech-1.5
- XTTS; https://huggingface.co/coqui/XTTS-v2
- F5-TTS; https://huggingface.co/ModelsLab/F5-tts-brazilian
- YourTTS; https://github.com/Edresson/YourTTS
- ToucanTTS; https://github.com/DigitalPhonetics/IMS-Toucan

ASR Model And Framework:

This model and framework were used to transcribe real and synthetic data in order to calculate the word error rate between models in dataset.

- Whisper Large V3 Turbo; https://huggingface.co/openai/whisper-large-v3-turbo
- Faster Whisper; https://github.com/SYSTRAN/faster-whisper

Metrics Tool:

- VERSA: Versatile Evaluation of Speech and Audio; https://github.com/wavlab-speech/versa