TAGARELA - A PORTUGUESE SPEECH DATASET FROM PODCASTS

Frederico Santos de Oliveira[⋆] *Lucas Rafael Stefanel Gris*[†] *Alef Iury Siqueira Ferreira*[†]
Augusto Seben da Rosa[‡] Alexandre Costa Ferro Filho[†] Edresson Casanova[§]
Christopher Dane Shulby[¶] Rafael Teixeira Sousa[⋆] Diogo Fernandes Costa Silva[†]
Anderson da Silva Soares[†] Arlindo Rodrigues Galvão Filho[†]

[⋆] Federal University of Mato Grosso (UFMT) [†] Federal University of Goias (UFG)
[‡] Paulista State University (UNESP) [§] NVIDIA [¶] Elsa Speak

ABSTRACT

Despite significant advances in speech processing, Portuguese remains under-resourced due to the scarcity of public, large-scale, and high-quality datasets. To address this gap, we present a new dataset, named TAGARELA, composed of over 8,972 hours of podcast audio, specifically curated for training automatic speech recognition (ASR) and text-to-speech (TTS) models. Notably, its scale rivals English's GigaSpeech (10kh), enabling state-of-the-art Portuguese models. To ensure data quality, the corpus was subjected to an audio preprocessing pipeline and subsequently transcribed using a mixed strategy: we applied ASR models that were previously trained on high-fidelity transcriptions generated by proprietary APIs, ensuring a high level of initial accuracy. Finally, to validate the effectiveness of this new resource, we present ASR and TTS models trained exclusively on our dataset and evaluate their performance, demonstrating its potential to drive the development of more robust and natural speech technologies for Portuguese. The dataset is released publicly to foster the development of robust speech technologies.

Index Terms— speech processing, text-to-speech, dataset, automatic-speech-recognition

1. INTRODUCTION

Portuguese ranks as one of the most widely spoken languages globally, with hundreds of millions of speakers across several continents. Despite this global prominence, it remains significantly under-resourced in the field of speech technology when compared to English. Recent advances in deep learning have propelled the fields of Automatic Speech Recognition (ASR) and Text-to-Speech (TTS), but their progress is fundamentally driven by the availability of large-scale, high-quality speech datasets. High-resource languages like English benefit from extensive corpora such as LibriSpeech (1000 hours) [1], GigaSpeech (10k hours) [2] and Emilia (139k hours) [3], which comprise tens of thousands of hours of transcribed audio. This disparity creates a significant bottleneck that hinders the development of robust and natural-sounding speech technologies tailored to the linguistic nuances of Portuguese.

To address this disparity, the research community has made commendable efforts to create and publicly release Portuguese speech corpora, with a notable focus on capturing the complexities of spontaneous speech. Landmark initiatives such as the CORAA corpus [4], which aggregates and manually validates 290 hours of spontaneous and prepared speech from various academic projects, have been pivotal. This trend has been further reinforced by the development of other valuable spontaneous speech datasets such as the NURC-SP Audio Corpus (239 hours) [5], the MuPe Life Stories Dataset (365 hours) [6], and the VoxCeleb-PT corpus (approx. 18 hours) [7], which focuses on celebrity voices for speaker recognition tasks. Although these corpora provide high-quality, manually revised data crucial for building more natural ASR systems, their inherent focus on spontaneous speech—often containing disfluencies, background noise, and interruptions—makes them ill-suited for training high-quality TTS models, which require pristine and consistent audio. Furthermore, they are orders of magnitude smaller than their English counterparts, limiting the performance ceiling of data-hungry state-of-the-art models.

A significant opportunity to bridge this data gap emerged with the release of the "Cem Mil Podcasts" collection [8], a massive corpus of Portuguese podcasts offering more than 76,000 hours of diverse, multi-dialect audio. However, this dataset presented a major barrier for a practical application: it was provided as raw, unprocessed audio with automatically generated transcripts of varying quality, which limits accessibility for new research. The lack of essential pre-processing, such as speaker diarization, noise reduction, and overlapping speech removal, rendered the dataset challenging to use directly for training high-performance ASR and TTS models, which require clean, single-speaker audio segments.

To unlock the potential of this resource, we introduce TAGARELA (/ta.ga.ˈrɛ.lɑ/), a large-scale Portuguese speech dataset curated from the "Cem Mil Podcasts" collection. We engineered a comprehensive pipeline—including audio

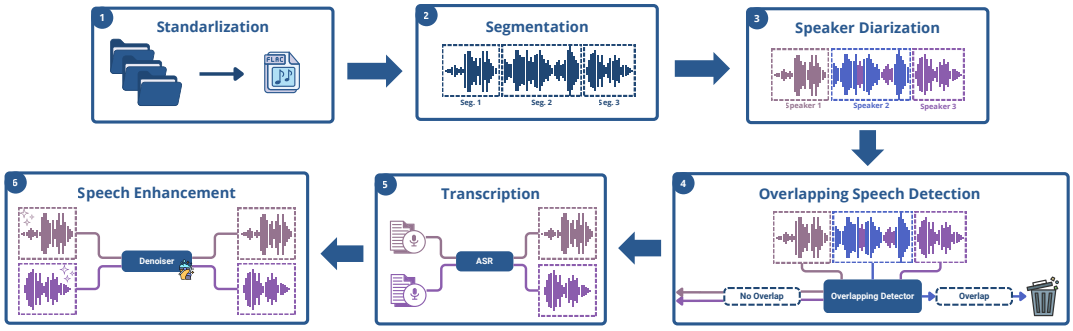

Fig. 1. Overview of the TAGARELA preprocessing pipeline.

standardization, speaker diarization, overlapping speech detection, and denoising—to transform the raw audio into a high-quality corpus. Transcription was accomplished via a scalable bootstrap method to balance quality and scalability: a 1000-hour "seed corpus", transcribed by a commercial ASR service, was used to fine-tune a Whisper large-v3 model. This specialized model, in turn, generated high-quality pseudo-labels for the final dataset. The resulting TAGARELA corpus is divided into two parts: a full subset of 8,972 hours, which includes audio containing various types of disfluencies for robust ASR training, and a clean-speech subset, 2,800-hour speech-only subset designed for speech generation tasks.

The main contributions of this work include the release of TAGARELA, a new curated Portuguese speech corpus with more than $8,972$ hours designed for ASR and TTS tasks, a detailed description and evaluation of our multi-stage data processing pipeline, and the training and evaluation of open-source models exclusively on the TAGARELA dataset, demonstrating its effectiveness.

2. TAGARELA DATASET

The TAGARELA dataset is a large-scale, Portuguese-language audio corpus from the "Cem Mil Podcasts" collection [8], created to enable research on information access and address the lack of non-English podcast resources. It consists of roughly 16,806 episodes from 2,094 shows, totaling over 8,972 hours of audio. The data includes both Brazilian (8,130 hours) and European Portuguese (842 hours) dialects. In terms of gender, 70% of the audio (6,368.34 hours) is attributed to male speakers and 30% (2,604.37 hours) to female speakers. The dataset's audio segments have an average duration of 9.30 ± 5.49 seconds and contain an average of 27.69 ± 17.06 words. Figure 2 presents the distribution of audio duration according to gender and accent.

3. TAGARELA PIPELINE

The creation of the TAGARELA dataset involved a multi-stage pipeline designed to ensure high quality and consistency

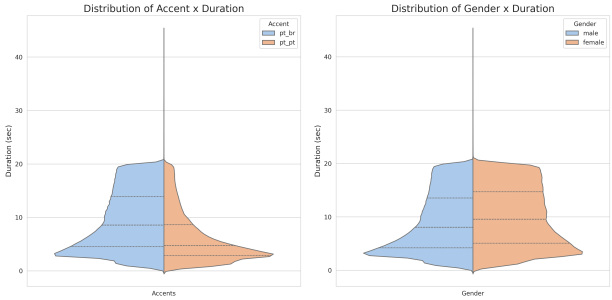

Fig. 2. Violin plots showing the distribution of audio segment duration in seconds. The left plot compares accents (pt_br vs. pt_pt), and the right plot compares genders (male vs. female).

for both ASR and TTS tasks. Each stage was planned to handle the challenges inherent to podcast audio, such as multiple speakers, background noise, and the need for accurate, large-scale transcriptions. Below, we detail each component of our pipeline.

3.1. Audio Standardization and Segmentation

Initially, all audio files in the corpus were processed to ensure a uniform format, an essential step for model training consistency. All audio was converted to the FLAC format, with a sample rate of 16kHz, 16-bit depth, and in a mono-channel. Following this, we segmented the long-form recordings into shorter clips. While the target duration was 5 to 20 seconds, the algorithm's primary goal was to split at natural silences. This approach maintains the cohesiveness of speech by preventing words or phrases from being cut off abruptly.

3.2. Diarization and Speaker Separation

A common feature in podcasts is the presence of multiple speakers, which poses a challenge for creating clean datasets. To address this, we applied a diarization process using the pyannote framework [9]. This stage identifies and labels the speech segments for each speaker individually. By separat-

ing the different speakers, we ensure that each final sample in the dataset contains the voice of only one person. This step is particularly crucial for training TTS models, which require single-speaker data to generate consistent voices.

3.3. Overlapping Speech Detection

Although diarization separates speakers, some segments may still contain overlapping speech where multiple speakers talk simultaneously. This is highly detrimental to the quality of TTS models. To mitigate this issue, we trained a dedicated classification model based on Wav2vec2-XLS-R[1] [10] to specifically identify these instances. All audio samples that were flagged by the model as containing overlapping speech were subsequently discarded from the dataset, ensuring that the final clips consist of clean, single-speaker utterances. To ensure the reproducibility of this filtering step, the trained model and its checkpoint are made publicly available for download.

3.4. Transcription Generation with a Mixed (Bootstrap) Strategy

We employed a bootstrap strategy for transcription. First, a high-fidelity "seed corpus" of approximately 1000 hours was generated with a commercial ASR service, ElevenLabs Scribe v1[2], to fine-tune a Whisper large-v3 model for pseudo-labeling. To filter Whisper's potential hallucinations [11], we also trained a Wav2vec2-XLS-R model[3] [12] on the same seed data. We then calculated the Word and Character Error Rates (WER/CER) between the outputs of both models, making it possible to select only samples with a high agreement score to ensure the training dataset's quality.

3.5. Quality Enhancement

For the final audio enhancement stage, we repurposed a Vocos vocoder [13] to act as a denoiser. The model was trained specifically for this task on a private dataset, optimizing it to remove common podcast artifacts like background noise, hiss, and light reverberation. This process significantly improves the clarity and quality of the finalized segments. To ensure the reproducibility of this pipeline, we also provide a version of the denoiser trained exclusively on public datasets.

3.6. Speaker and Dialect Labeling

To enrich the dataset, we implemented a multi-stage labeling process. First, given that the original data lacks speaker labels, we performed a cluster-based speaker labeling step. We extracted embeddings for each audio segment using the RedimNet B6 model [14] and grouped them with the HDBSCAN algorithm [15]. This process was performed independently for each podcast to avoid merging speaker identities, resulting in approximately 13,368 distinct speaker labels.

Furthermore, we developed a model to classify each segment's dialect as either Brazilian or European Portuguese. This was achieved by first pre-training a wav2vec-base model on all segmented audio from our dataset. Subsequently, this model was fine-tuned on a balanced combination of the CORAA, CommonVoice, and CML-TTS datasets to create the final accent classifier thus adding valuable dialectal metadata to the corpus.

4. EXPERIMENTS AND EVALUATION

In this section, we validate the quality and effectiveness of the TAGARELA dataset by using it to train state-of-the-art models for ASR and TTS. Our goal is to demonstrate that the data curated through our pipeline can produce models with competitive or state-of-the-art performance for Portuguese.

4.1. Objective Metrics

We assess audio quality using three objective metrics: Short-Time Objective Intelligibility (STOI) [16] for speech intelligibility, the wideband version of Perceptual Evaluation of Speech Quality (PESQ) [17, 18] for perceived quality, and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [19] for signal fidelity in decibels (dB). For all metrics, higher values indicate better quality. Since these traditionally require a clean reference signal, we employ the TorchAudio-Squim [20] framework to obtain reference-free estimates. The results are presented in Figure 3.

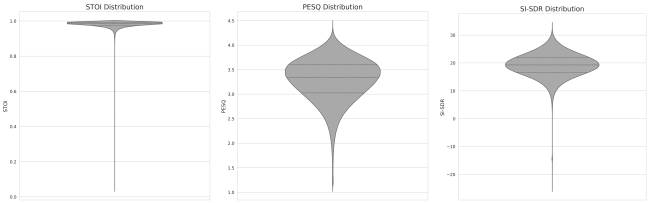

Fig. 3. Violin plots showing STOI, PESQ and SI-SDR.

4.2. Speech Recognition (ASR) Experiments

To evaluate the potential of TAGARELA for ASR tasks, we trained a diverse set of model architectures, covering different sizes and capabilities: Parakeet TDT [21, 22], Canary-1b-Flash and Distil-Whisper [23].

Training Setup: All models were trained through fine-tuning using a full 8,972 hour subset of the TAGARELA dataset. Training was conducted using either an NVIDIA A100 or B200 GPU, subject to availability in our experiment runtime.

[1]https://huggingface.co/facebook/wav2vec2-xls-r-300m
[2]https://elevenlabs.io/docs/models#scribe-v1, run in June 2025.
[3]https://huggingface.co/facebook/wav2vec2-xls-r-1b

Evaluation: We evaluated model performance on the Portuguese (pt) split of the Common Voice 17.0[4] dataset, using Word Error Rate (WER)—calculated after text normalization—as the primary metric.

Results: As detailed in Table 1, our evaluation on the Common Voice 17.0 test set highlights the strong performance of Canary-1B-Flash. The model achieved the lowest WER of 7.8%, outperforming both Distil-Whisper (9.2%) and Parakeet TDT (12.3%). These results position Canary-1B-Flash as a highly effective and efficient model for Portuguese speech recognition.

Table 1. WER results on the Common Voice 17.0 (pt) test set.

Model	WER (%) ↓
Canary-1B-Flash	**7.8**
Distil-Whisper	9.2
Parakeet TDT	12.3

4.3. Text-to-Speech (TTS) Experiments

Training and Evaluation Setup: For the TTS task, we trained the Orpheus-TTS[5] and Chatterbox[6] models using the 2,800-hour clean-speech subset of the TAGARELA dataset. We evaluated performance on two fronts: *intelligibility*, measured by WER/CER using Whisper Large V3, and *perceptual quality*, assessed with a Mean Opinion Score (MOS) from the SHEET benchmark[7] main model. To ensure a robust evaluation, we applied a two-stage outlier removal process, filtering samples shorter than five seconds, and applying a quartile-based method to remove statistical outliers.

Table 2. TTS model performance on the CML. The values for CER, WER, and MOS are presented as mean ± standard deviation.

Model	CER (%) ↓	WER (%) ↓	MOS ↑
Chatterbox	23.73 ± 26.17	31.50 ± 30.05	$\mathbf{4.53} \pm 0.25$
Orpheus-TTS	$\mathbf{19.32} \pm 31.64$	$\mathbf{26.81} \pm 35.57$	4.00 ± 0.94

Results and Analysis: As shown in Table 2, Orpheus-TTS delivered superior intelligibility with a WER of 26.81%, while Chatterbox achieved a more natural-sounding voice with a MOS of 4.53, despite its higher error rates. This suggests that Orpheus-TTS excels at clarity, while Chatterbox prioritizes perceived quality.

These experiments validate the TAGARELA dataset's critical role in advancing Portuguese TTS. Despite the dataset's imperfect text-audio alignment, the results are highly encouraging and provide a solid foundation for developing robust, high-quality TTS systems for Portuguese.

[4]https://huggingface.co/datasets/mozilla-foundation/common_voice

[5]https://github.com/canopyai/Orpheus-TTS

[6]https://github.com/resemble-ai/chatterbox

[7]https://github.com/unilight/sheet

5. CONCLUSION

To address the resource gap in Portuguese speech technology, we introduce TAGARELA, a new large-scale dataset with over 8,972 hours of podcast audio. We presented a comprehensive pipeline using diarization, denoising, and a scalable transcription strategy to create a high-quality corpus suitable for both ASR and TTS. The public release of this dataset is a significant contribution, offering the community a resource on a scale previously unavailable for the Portuguese language.

The effectiveness of TAGARELA was validated by training ASR and TTS models exclusively on our data, achieving highly competitive performance. This confirms the dataset's potential to drive significant advancements in Portuguese speech processing. While there is room for refinements, such as improving text-audio alignment, TAGARELA offers a robust foundation for future innovations. We believe this resource will foster the development of more accurate and natural speech technologies, benefiting millions of Portuguese speakers.

Acknowledgements: 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 of the MCTI grant number 057/2023, signed with EMBRAPII.

6. REFERENCES

[1] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[2] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu, "Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 885–890.

[3] Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Interspeech 2021*. Aug. 2021, interspeech_2021, ISCA.

[4] Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra Maria Aluísio, "Coraa asr: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese," *Language Resources and Evaluation*, vol. 57, no. 3, pp. 1139–1171, 2023.

[5] Rodrigo Lima, Sidney Evaldo Leal, Arnaldo Candido Junior, and Sandra Maria Aluisio, "A large dataset of spontaneous speech with the accent spoken in são paulo for automatic speech recognition evaluation," in *Proceedings of 34th Brazilian Conference on Intelligent Systems (BRACIS)*, 2024.

[6] "MuPe life stories dataset: Spontaneous speech in Brazilian Portuguese with a case study evaluation on ASR bias against speakers groups and topic modeling," in *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, Eds., Abu Dhabi, UAE, Jan. 2025, pp. 6076–6087, Association for Computational Linguistics.

[7] John Mendonça and Isabel Trancoso, "Voxceleb-pt–a dataset for a speech processing course," *Proc. IberSPEECH*, vol. 2022, pp. 71–75, 2022.

[8] Ekaterina Garmash, Edgar Tanaka, Ann Clifton, Joana Correia, Sharmistha Jat, Winstead Zhu, Rosie Jones, and Jussi Karlgren, "Cem mil podcasts: A spoken portuguese document corpus for multi-modal, multi-lingual and multi-dialect information access research," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Cham, 2023, pp. 48–59, Springer Nature Switzerland.

[9] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "Pyannote.audio: Neural building blocks for speaker diarization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7124–7128.

[10] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[11] Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk, "Investigation of whisper asr hallucinations induced by non-speech audio," 04 2025, pp. 1–5.

[12] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[13] Hubert Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," in *International Conference on Representation Learning*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, Eds., 2024, vol. 2024, pp. 25719–25733.

[14] Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov, "Reshape dimensions network for speaker recognition," in *Interspeech 2024*, 2024, pp. 3235–3239.

[15] Claudia Malzer and Marcus Baum, "A hybrid approach to hierarchical density-based cluster selection," in *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2020, pp. 223–228.

[16] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[17] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[18] I. Rec, "P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Recommendation P.862.2, International Telecommunication Union, Geneva, Switzerland, 2005.

[19] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[20] Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg, "Fast conformer with linearly scalable attention for efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[22] Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg, "Efficient sequence transduction by jointly predicting tokens and durations," in *Proceedings of the 40th International Conference on Machine Learning*. 2023, ICML'23, JMLR.org.

[23] Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush, "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling," *arXiv preprint arXiv:2311.00430*, 2023.