FLA-Dataset: A Database of Four Location-Based Audio Commands in Brazilian Portuguese

Daniel Ribeiro da Silva[1], **Gabriel Pettro Oliveira Ruotolo**[2], **Alexandre Costa Ferro Filho**[3], **Marcelo Henrique Lopes Ferreira**[3], **Letícia Lima Mendes**[1], **José Rafael Rebêlo Teles**[1]

[1]Institute of Informatics
Federal University of Goiás
Goiânia – GO – Brazil

[2]School of Electrical, Mechanical, and Computer Engineering
Federal University of Goiás
Goiânia – GO – Brazil

[3]Advanced Knowledge Center in Immersive Technologies (AKCIT)
Federal University of Goiás
Goiânia – GO – Brazil

{daniel.ribeiro,gabrielruotolo,alexandre_ferro,limamendes,

joserafael}@discente.ufg.br, marcelomarcelo2@egresso.ufg.br

Abstract. *The FLA-Dataset provides a curated collection of wake word audio commands in Brazilian Portuguese. It supports the development and evaluation of voice-activated systems by including four direction-oriented commands: direita, esquerda, frente, and pare. Recordings are contributed by 30 speakers across diverse demographic and acoustic profiles. The dataset promotes generalization by incorporating varied speech patterns, environments, and genders. All files are standardized in format, sampling rate, and structure to ensure usability. Each sample contains a single spoken command and is organized to allow speaker-specific experiments. Applications span robotics, assistive technologies, and smart devices. The dataset addresses the scarcity of localized resources for Portuguese. It enables speaker-independent training and evaluation in realistic scenarios. FLA-Dataset is publicly available and designed to support inclusive speech recognition research and deployment.*

1. Introduction

Wake word detection is a fundamental component in the operation of voice-activated systems. It enables devices to remain in a low-power listening state until a specific keyword is spoken, thereby initiating active processing and interaction. This mechanism is widely applied in virtual assistants, smart home devices, and mobile applications, enhancing both usability and energy efficiency [Warden 2018]. Wake words serve as natural, intuitive triggers for human-computer interaction, supporting seamless and hands-free control in environments where manual input may be impractical or impossible [Lim et al. 2023].

The development of effective wake word detection models strongly depends on the availability of robust and diverse datasets [Ribeiro et al. 2023]. High-quality datasets al-

low for better generalization of models across acoustic environments, speaker identities, and microphone characteristics. The presence of noise, speaker variation, and channel mismatch are critical factors addressed by well-curated datasets. Furthermore, as deep learning approaches continue to evolve, the demand for large-scale and accurately labeled audio data for wake word detection becomes increasingly vital for both research and commercial applications [Chen et al. 2014].

There remains a scarcity of datasets tailored to Brazilian Portuguese that also encompass a significant degree of speaker variability [Stefanel Gris et al. 2022]. Most existing public wake word datasets are concentrated in English, limiting the development of voice-interaction systems for non-English-speaking populations. Additionally, datasets in Portuguese often feature limited diversity in accents and speaker demographics, which constrains the performance and adaptability of wake word models in real-world applications. This gap highlights the need for localized datasets that reflect the linguistic and acoustic diversity of Brazilian Portuguese speakers across different regions and contexts.

Therefore, this work proposes the FLA-Dataset, a wake word dataset in Brazilian Portuguese designed to support the development of voice activated systems. The dataset includes four activation words: *direita*, *esquerda*, *frente*, and *pare*. These words correspond to common control and directional commands used in areas such as robotics, intelligent assistants, and accessibility focused technologies. The dataset emphasizes speaker diversity, incorporating samples from a wide range of voices, regions, and speaking styles. This diversity enhances the training and evaluation of systems in real world conditions. The chosen words span different phonetic structures and syllabic patterns, introducing acoustic variety that challenges recognition models and encourages the development of more accurate and robust speech recognition systems. By addressing the lack of publicly available resources in Portuguese with this focus, the FLA-Dataset fills an essential gap in the research and application of spoken command systems.

The objective is to construct a comprehensive dataset that can support further research in wake word detection by offering a solid foundation in Brazilian Portuguese. This dataset aims to foster advancements in model evaluation, comparison, and development, addressing both academic and applied research needs. It seeks to serve as a benchmark resource for the evaluation of wake word models in scenarios requiring high speaker variability and contextual relevance in Portuguese language settings. The motivation for creating the dataset stems from the need to test a complete pipeline for wake word creation and deployment in a robotic context, as proposed in earlier work [Ferro Filho et al. 2025]. The dataset introduced in this study focuses exclusively on real audio recordings, providing a reliable testing resource that enables consistent and meaningful assessment of wake word detection models under natural acoustic and speaker conditions.

The originality of the work lies in the combination of linguistic localization, phonetic diversity, and broad speaker representation within the context of wake word detection. Few publicly available datasets integrate these three aspects in the Portuguese language. This research contributes to filling this gap and supports the development of inclusive and adaptable speech technologies. Its relevance extends to areas such as assistive technology, human-robot interaction, and smart device integration, where efficient and accurate wake word detection in diverse linguistic settings remains a technical challenge [Ribeiro et al. 2023].

2. Related Works

The Speech Commands dataset [Warden 2018] provides a standardized benchmark for limited-vocabulary speech recognition and serves as a foundational resource for developing wake word detection systems. It contains over 100,000 audio recordings of 35 isolated words recorded by more than 2,600 speakers, sampled at 16 kHz. The dataset includes both command-related keywords and distractors, along with background noise and silence, enhancing robustness for real-world scenarios. Recordings were collected via web browsers using volunteers' microphones, ensuring acoustic and pronunciation diversity. The dataset includes official splits for training, validation, and testing, which supports reproducibility and comparability across models. Its permissive license and broad adoption have enabled advancements in keyword spotting across different neural architectures. However, it focuses exclusively on English, which limits its applicability for multilingual and culturally diverse speech interfaces.

To address the need for linguistic diversity in wake word detection, the Robat-e-Beheshti dataset [Raji and Shekofteh 2022] extends dataset availability to the Persian language, with an emphasis on applications in robotics. It includes 5,738 audio files of up to 3 seconds each, recorded by 187 individuals using various devices. The dataset provides both positive and negative samples and standardizes all files in terms of format and sampling rate. To increase variability, it integrates samples from the ShEMO emotional speech dataset. This dataset supports research in Persian speech command recognition and fosters linguistic inclusion in human-robot interaction. Its development reflects the complexity of ensuring speaker diversity, recording quality, and class balance in underrepresented languages, emphasizing the challenges of building reliable wake word datasets in low-resource scenarios.

Beyond dataset creation, various strategies have been proposed to mitigate data scarcity. The study by Ribeiro et al. [Ribeiro et al. 2023] explores training strategies that minimize reliance on large annotated datasets for wake word detection. It compares alignment-based, alignment-free, and hybrid methods. The alignment-based approach uses phonetic annotations with cross-entropy loss, while the alignment-free method applies Connectionist Temporal Classification to avoid phonetic alignment. The hybrid strategy combines both, using a small aligned subset to guide the training of a larger unaligned set. Results show that the alignment-free model is more effective at reducing false activations, while the hybrid approach balances accuracy and data efficiency. The study demonstrates that training competitive models is feasible using only ten to twenty percent of aligned data, highlighting efficient alternatives when large annotated datasets are not accessible.

Further advances in data efficiency have focused on model adaptation techniques for constrained environments. Cioflan et al.[Cioflan et al. 2024] introduce a domain adaptation strategy for low-power devices that fine-tunes only the final classification layer of keyword spotting models. This adaptation requires only one hundred labeled samples and less than ten kilobytes of memory, completing in under fifteen seconds. The method enhances accuracy by up to fourteen percent in noisy conditions, supporting privacy-preserving and energy-efficient applications in TinyML and IoT environments. Similarly, Lim et al.[Lim et al. 2023] propose LiteFEW, a lightweight encoder based on wav2vec 2.0 that discards transformer layers in favor of convolutional layers combined with autoen-

coding and knowledge distillation. The encoder remains frozen, while only the detection model is updated. Evaluated on the Hey Snips dataset, LiteFEW achieves high accuracy while reducing parameters by up to ninety-nine percent, making it suitable for wake word detection on devices with limited computational capacity and access to small datasets.

The use of synthetic and weakly labeled data further expands the possibilities for training robust models under data constraints. Wang et al.[Wang et al. 2020] present a wake word detection method based on alignment-free lattice-free maximum mutual information, allowing training with partially labeled data. The approach uses a simplified HMM topology, a compact acoustic model, and extensive data augmentation such as noise addition, reverberation, and speed perturbation. These augmentations increase the training set size by seven times, enhancing robustness and generalization. The system maintains high performance across various datasets with efficient online decoding and a lightweight architecture, validating the benefits of synthetic data and alignment-free training. Complementing this direction, Coucke et al.[Coucke et al. 2018] propose the Snips Voice Platform, a privacy-focused system for embedded spoken language understanding. To ensure user privacy, the system avoids cloud-based training and relies on synthetic data and crowdsourcing. It generates diverse training samples through simulated environments, artificial noise, and reverberation, enabling the development of compact models that run entirely on-device.

Despite significant progress in reducing reliance on large annotated datasets, diverse and representative data remain crucial for building robust wake word detection systems. Approaches such as alignment-free training, domain adaptation, and synthetic data generation improve performance under data constraints, but their effectiveness depends on the quality and diversity of available samples. Multilingual and culturally specific datasets are essential for inclusive voice interfaces that generalize across speakers, accents, and environments. However, there is still a lack of publicly available wake word datasets in Brazilian Portuguese, which limits research and development for voice-based systems in this language. Expanding dataset availability to underrepresented languages strengthens benchmarking, reproducibility, and the advancement of keyword spotting across domains.

3. Methodology

This section presents the methodology employed for the creation of the FLA-Dataset, detailing the complete pipeline from data acquisition to dataset publication. The process includes the selection and recording of activation words, audio preprocessing procedures, manual verification to ensure the quality and integrity of the samples, and the final structuring and organization of the dataset. Each step was designed to support the development of robust and generalizable wake word detection models. Figure 1 illustrates the overall workflow, highlighting the key stages involved in the dataset construction.

3.1. Data Acquisition

The selection of the activation words focuses on practical commands commonly employed in robotic and voice-controlled systems. The words *direita*, *esquerda*, *frente*, and *pare* allow for clear and direct control over directional actions, especially in navigation-based scenarios such as human-robot interaction, assistive technologies, and autonomous mobility platforms. These words are phonetically distinct, which benefits acoustic separation and recognition by speech models. Their use also aligns with typical command

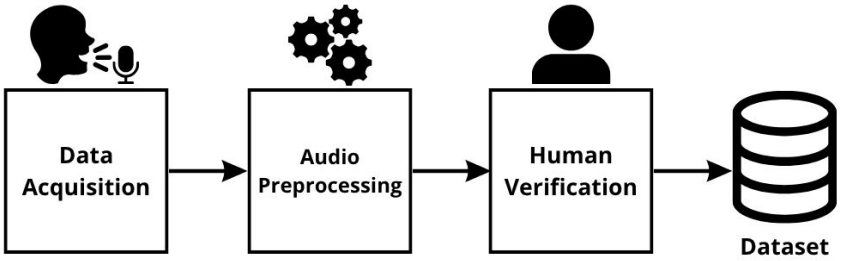

Figure 1. General overview of the methodology.

structures in Brazilian Portuguese, which contributes to natural language integration in control systems.

Recordings are conducted with a diverse group of speakers encompassing a wide range of vocal characteristics such as pitch, tone, accent, and articulation. Participants capture the audio primarily using their own mobile devices, which introduces natural variation in terms of recording environments and microphone types. This diversity contributes to the realism of the dataset and exposes models to different acoustic conditions without the need for artificial simulation. However, the dataset does not include explicit labels or metadata indicating the specific devices or environments used in each recording. All individuals involved in the recordings receive specific training to adapt to the technical and procedural requirements of the audio capture platform. This preparation aims to ensure consistency in pronunciation, microphone handling, and recording settings while maintaining the individuality of each speaker's vocal identity.

The audio collection process followed a standardized protocol where each speaker recorded at least thirty seconds of speech for each of the four wake words. Each utterance of the wake word was separated by a short interval of silence to facilitate segmentation during preprocessing. The repetition of the words over an extended period allowed the capture of natural variations in speech time and intonation. All target words are represented by a diverse range of speakers, resulting in a well-balanced dataset. This design supports a wide range of experiments, including speaker-independent and noise-robust wake word detection tasks.

3.2. Audio Preprocessing

The preprocessing stage began with the segmentation of the original audio recordings. Each recording, which initially contained several repetitions of a wake word spoken by a single speaker, was manually segmented by an annotator. This process ensured that each resulting audio file contained only one utterance of a specific wake word. The segmentation respected the identity of each speaker, maintaining the association between the extracted audio segments and their respective speaker labels. This structure preserves the integrity of speaker-based analyses and allows for the development of models that consider speaker variability.

To ensure consistency across the dataset, all audio files were standardized to the same format. Each file was converted to a single channel and resampled to a frequency of 16 kilohertz (16 KHz), a common configuration in speech processing tasks

[Panayotov et al. 2015]. This standardization allows models to process data uniformly and supports reproducibility in experiments. In addition to the sampling rate, all audio files were stored in WAV format to maintain compatibility with most speech recognition frameworks and ensure lossless audio quality.

3.3. Final Dataset and Usability

The final version of the dataset is structured into four main directories, each corresponding to one of the activation words: *direita*, *esquerda*, *frente*, and *pare*. Within each directory, audio files are named following a consistent format that facilitates traceability and organization. The naming convention follows the pattern <word>_s<speaker_id>_a<sample_id>.wav, where sXXX identifies the speaker and aYYY represents the sample number. This organization enables easy parsing and filtering of files for custom training and evaluation pipelines, especially when designing experiments that rely on speaker-specific data.

The speaker distribution is defined to support various types of demographic analyses. A total of 30 speakers contribute to the dataset, including 10 female speakers with identifiers ranging from s000 to s009 and 20 male speakers with identifiers from s010 to s029. This balance enables experiments that consider gender-based variability, speaker-dependent and speaker-independent training strategies, and evaluations focused on model generalization. The dataset allows for stratified data splits based on speaker ID or gender, which enhances its flexibility and relevance in multiple research scenarios.

The FLA-Dataset is publicly available on the Zenodo platform at *https://doi.org/10.5281/zenodo.15677186*, and is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0) [1]. The open and accessible license promotes reproducibility and supports research in Brazilian Portuguese speech technologies. The dataset maintains high phonetic quality, balanced representation across speakers, and standardized recording conditions. These characteristics ensure reliability for experimental validation, facilitate comparative evaluations among models, and contribute to the advancement of speech processing in diverse environments.

4. Exploratory Analysis

The dataset contains recordings from 30 speakers, including 10 female and 20 male individuals. The total duration of all recordings reaches 40.78 minutes, and the average duration per audio file is 1.05 seconds. Each recording consists of a single spoken word, ensuring consistency and clarity across the dataset. The spoken words are *direita*, *esquerda*, *frente*, and *pare*, with 624, 514, 597, and 594 samples respectively. This balanced distribution across classes supports fair representation and enhances the reliability and generalization potential of models developed using this data.

Figure 2 (A) illustrates the distribution of spoken words by speaker gender. Although male speakers contribute approximately twice as many samples as female speakers, the difference does not compromise the representativeness of both groups. Each command word receives contributions from speakers of both genders in all cases, preserving diversity across classes. The presence of both male and female voices across all categories

[1] https://creativecommons.org/licenses/by/4.0/

continues to support the development of models that generalize well and perform reliably across different user profiles. This structure helps reduce the risk of gender-related performance variation and strengthens the robustness of speaker-independent systems.

Figure 2 (B) shows the number of samples per speaker. The dataset includes a significant number of speakers, allowing a diverse representation of voice characteristics and speaking styles. This diversity contributes to a richer and more realistic audio collection, suitable for studying speaker variability in wake word detection. Each speaker provides multiple recordings, enabling the analysis of consistency and robustness of models under different vocal conditions. The separation of speakers within the dataset also allows investigations on generalization, making it possible to evaluate model performance when trained on specific individuals and tested on others. This structure enhances the potential for research focused on speaker adaptation, speaker-independent recognition, and cross-speaker behavior in voice-controlled systems.

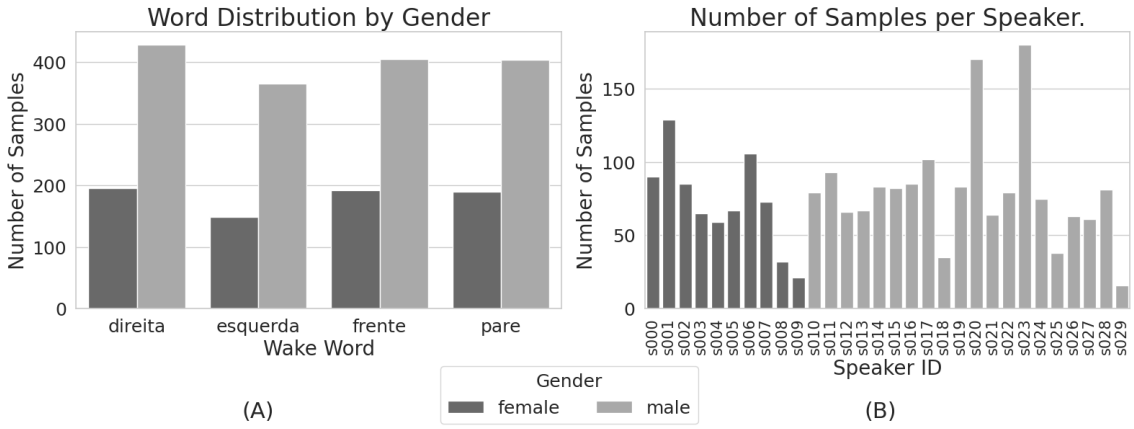

Figure 2. FLA-Dataset Statistics: (A) Number of audio samples per wake word by gender, and (B) Number of samples per speaker by gender.

For a simple benchmark, the FLA-Dataset was used to evaluate the OpenWakeWord [2] framework, which trains models using synthetic speech data. Table 1 reports the accuracy, recall, and F1-score metrics for each keyword. The overall results demonstrate that the dataset poses a challenging scenario for wake word detection, primarily due to the short duration of the commands, acoustic variability among synthetic voices, and phonetic similarity between keywords (e.g., “pare” and “frente”). These factors lead to a notable drop in recall, indicating that the models often fail to trigger correctly in the presence of subtle prosodic or temporal variations. Such behavior confirms that the FLA-Dataset provides a reliable benchmark for assessing model robustness and generalization, especially under conditions that mimic real-world speech variability in human–machine interaction systems.

5. Applications

Wake word detection plays a fundamental role in enabling natural and efficient voice-based interaction across diverse application domains such as smart assistants, autonomous vehicles, wearable technologies, and robotic platforms. Its primary function is to activate

[2]https://github.com/dscripka/openWakeWord

Table 1. OpenWakeWord Performance on the FLA-Dataset.

Keyword	Accuracy	Recall	F1
Direita	0.78	0.17	0.58
Esquerda	0.89	0.52	0.81
Frente	0.76	0.07	0.50
Pare	0.75	0.01	0.43

voice interfaces by continuously monitoring for predefined keywords while using minimal computational resources. This strategy promotes energy efficiency and enhances user experience. Wake word systems are especially useful in scenarios where manual control is impractical or undesirable, including assistive technologies for users with motor impairments and robotic systems that operate in dynamic or time-sensitive environments.

In this context, the FLA-Dataset provides a specialized resource designed for Brazilian Portuguese, a language variety with limited representation in publicly available datasets. The dataset focuses on four practical and context-relevant commands - *direita*, *esquerda*, *frente*, and *pare* - selected for their alignment with navigation and control tasks. These commands support a variety of voice-driven applications across key areas, including:

- **Human-Robot Interaction (HRI)**: Voice commands for navigation and control are essential in robotic systems for tasks such as teleoperation, exploration, and safety procedures like emergency stop. The FLA-Dataset offers real recordings from a diverse set of speakers, contributing to the development of models that generalize well in robotic environments.
- **Assistive Technologies**: Spatial commands contained in the dataset help enable systems that offer greater autonomy to users with physical limitations by supporting natural interaction through voice alone.
- **Smart Home and IoT Devices**: The directional nature of the commands aligns with frequent use cases in smart appliances and domestic robotics, improving the integration of voice control in Portuguese-speaking environments.
- **Embedded and Edge AI Systems**: Models trained using the FLA-Dataset can operate on low-power hardware for offline voice activation, supporting the principles of TinyML while ensuring user privacy and fast local response.

By addressing the scarcity of Portuguese-language resources for wake word detection, the FLA-Dataset contributes to the development of inclusive voice technologies and strengthens the foundation for speech-based interfaces in assistive, embedded, and robotic applications. Its relevance extends to both academic research and real-world deployment, particularly in linguistic contexts that remain underrepresented in speech technology development.

6. Conclusion and Future Works

The FLA-Dataset addresses a significant gap in the development of wake word detection systems adapted to Brazilian Portuguese. Its design emphasizes phonetic diversity, speaker balance, and practical relevance, making it a valuable resource for evaluating speech recognition models. The inclusion of four direction-oriented commands allows

for integration into areas such as robotics, assistive technologies, and embedded systems. The dataset's structure allows for speaker-independent experiments, gender variability, and acoustic generalization, contributing to more inclusive and robust voice interaction technologies. The dataset's main limitation is its limited vocabulary, consisting of only four wake words, which reduces its applicability in scenarios that require greater command diversity or linguistic variation.

Future developments may focus on augmenting the dataset with additional activation words and negative samples to support broader classification tasks and false positive mitigation. The integration of recordings captured in various noisy settings would also increase model robustness. Expanding demographic diversity among speakers and including emotion-laden or spontaneous speech could further improve the dataset's representativeness. The creation of benchmarking tasks, such as predefined splits and baseline models, would support standardized comparisons across systems. These extensions aim to strengthen the dataset's role as a foundational resource in the evolution of speech-driven systems in Brazilian Portuguese.

Acknowledgments

The authors would like to thank the Núcleo de Robótica Pequi Mecânico for their support in the development of the dataset described in this work. The technical and structural assistance provided by the group was essential for the data collection, organization, and validation processes, contributing significantly to the quality and feasibility of the proposed study.

References

[Chen et al. 2014] Chen, G., Parada, C., and Heigold, G. (2014). Small-footprint keyword spotting using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4087–4091.

[Cioflan et al. 2024] Cioflan, C., Cavigelli, L., Rusci, M., de Prado, M., and Benini, L. (2024). On-device domain learning for keyword spotting on low-power extreme edge embedded systems. In *2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS)*, pages 6–10.

[Coucke et al. 2018] Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., Primet, M., and Dureau, J. (2018). Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint*, arXiv:1805.10190.

[Ferro Filho et al. 2025] Ferro Filho, A. C., Ribeiro da Silva, D., Rebêlo Teles, J. R., Pettro Ruotolo, G., Mendes, L., Lopes Ferreira, M. H., and Woerle de Lima Soares, T. (2025). A simplified pipeline for wakeword creation and deployment: Leveraging zero-shot text-to-speech and ros2 for robotic systems. In *2025 Brazilian Conference on Robotics (CROS)*, volume 1, pages 1–6.

[Lim et al. 2023] Lim, H., Kim, Y., Yeom, K., Seo, E., Lee, H., Choi, S. J., and Lee, H. (2023). Lightweight feature encoder for wake-up word detection based on self-supervised speech representation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

[Panayotov et al. 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210.

[Raji and Shekofteh 2022] Raji, P. A. and Shekofteh, Y. (2022). Robat-e-beheshti: A persian wake word detection dataset for robotic purposes. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 434–439.

[Ribeiro et al. 2023] Ribeiro, V., Huang, Y., Shangguan, Y., Yang, Z., Wan, L., and Sun, M. (2023). Handling the alignment for wake word detection: A comparison between alignment-based, alignment-free and hybrid approaches. In *Interspeech 2023*, pages 5366–5370.

[Stefanel Gris et al. 2022] Stefanel Gris, L. R., Casanova, E., de Oliveira, F. S., da Silva Soares, A., and Candido Junior, A. (2022). Brazilian portuguese speech recognition using wav2vec 2.0. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022*, page 333–343.

[Wang et al. 2020] Wang, Y., Lv, H., Povey, D., Xie, L., and Khudanpur, S. (2020). Wake word detection with alignment-free lattice-free mmi. In *Interspeech 2020*, pages 4258–4262.

[Warden 2018] Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. *ArXiv preprint*, arXiv:1804.03209.