



SiDi KWS: A Large-Scale Multilingual Dataset for Keyword Spotting

Michel Meneses, Rafael Holanda, Luis Peres, Gabriela Rocha

SiDi, Brazil

{m.meneses, r.holanda, l.peres, g.rocha}@sidi.org.br

Abstract

Keyword spotting (KWS) has become a hot topic in speech processing due to the rise of commercial applications based on voice command detection, such as voice assistants. Like tasks in computer vision, natural language processing, and even speech processing, most current successful approaches for KWS rely on deep learning. However, differently from all those tasks, there is a lack of large-scale datasets designed for training and evaluating deep learning models for KWS. The current work presents SiDi KWS, a public large-scale multilingual dataset currently composed of 24.3 million audio recordings of labeled single-spoken keywords. It intends to boost the development of new KWS systems, especially those based on deep learning. That dataset has been created by applying automatic forced alignment on public datasets of transcribed speech. This work introduces SiDi KWS and KeywordMiner, an open-source framework used to generate that dataset, to benefit the speech processing research community.

Index Terms: keyword spotting, KWS, speech processing, dataset, multilingual

1. Introduction

Keyword spotting (KWS) corresponds to the task of detecting a target spoken keyword in speech utterances [1]. It is a core component of many speech processing applications, such as spoken document retrieval [2, 3] and voice control systems [4, 5, 6]. The latest include conversational agents (*e.g.*, Siri, Alexa, Bixby), which have been incorporated into many smart devices over the past decade. Those agents rely on KWS to trigger their core functionalities once their predefined wake-up keyword is spoken by the user. KWS, therefore, plays a vital role in the user experience for those applications.

Due to the relevance of KWS for commercial applications based on voice interfaces, that task has received a lot of attention from the speech processing and machine learning communities in the past few years [7, 8, 9]. That emphasis is in part boosted by the unique constraints usually involved in KWS applications, mainly the focus on detecting only a small set of target keywords and the need of performing fewer computations [10]. The latest rises from the use of KWS in low-power edge devices in an always-listening mode [11]. Modern approaches for KWS apply a single classifier on spectral representations of audio segments extracted from the input stream (*e.g.*, log-Mel spectrograms, Mel-frequency cepstral coefficients). Since the rise of deep learning in the previous decade, classifiers based on deep neural networks (DNN) have become the most successful approaches for solving KWS [12, 13, 14, 6]. More specifically, convolution neural networks (CNN) have been successful in solving that task. By sharing weights in their input space, those networks can model their spectral input's local correlations in time and frequency. At the same time, sharing weights reduces the number of parameters necessary for the network to

become invariant to input translations when compared to fully-connected networks [13]. In fact, KWS is only one of many tasks that benefit from CNN's architectures, including tasks from fields other than speech processing, such as computer vision [15], weather forecasting [16] and biomedicine [17].

Due to their data-hungry nature, deep learning models demand a large number of training samples to solve machine learning tasks. The more training samples are provided to a deep learning model and the more diverse those samples are, the lower tends to be its training and testing errors [18]. The recent rise of deep learning is partially justified by the high availability of structured and formatted data [19, 20, 21]. However, collecting, organizing, and formatting large datasets from scratch is a laborious task. That is why large public datasets play a vital role in the advance of deep learning research, as observed in the fields of computer vision [22, 23], natural language processing [24, 25], and speech processing [26, 27]. By having access to those datasets, researchers can focus their effort on designing better learning algorithms without worrying about collecting new data. Additionally, it becomes easier to benchmark algorithms and track how a particular research field advances. Finally, models trained on those datasets can be used as the starting point for developing commercial models, given the use of proper transfer learning techniques [28].

Although KWS research shares many of the deep learning approaches observed in other tasks from different fields (*e.g.*, CNN [13], residual blocks [29], and attention mechanisms [30]), it still lacks large public datasets. Differently from the large datasets publicly available for tasks such as image classification [22], object detection [23], and even automatic speech recognition [26], the majority of public datasets for KWS currently available are small and have low diversity, either in terms of vocabulary [31] as well as in languages [10] and speakers [32]. This work introduces SiDi KWS, a large-scale multilingual public dataset of labeled single-spoken keywords. That is the current largest public KWS dataset in terms of the number of audio samples and has a large vocabulary. The recordings that compose that dataset were extracted from popular public transcribed speech datasets (*i.e.*, LibriSpeech [26], Mozilla Common Voice [27] and MultiLingual LibriSpeech (MLS) [33]). Given the original speech audio files and their transcriptions, the segmentation of the single keywords was performed via forced alignment, by using the public tool Montreal Forced Aligner [34]. The process of loading those transcribed speech datasets, performing the forced alignment, segmenting the keywords, and exporting their labeled audio files was implemented during this work as an open-source framework named Keyword Miner. Overall, this work has the following goals:

- Introduce SiDi KWS, the largest public dataset for keyword spotting.
- Present Keyword Miner, the open-source framework used to generate SiDi KWS.

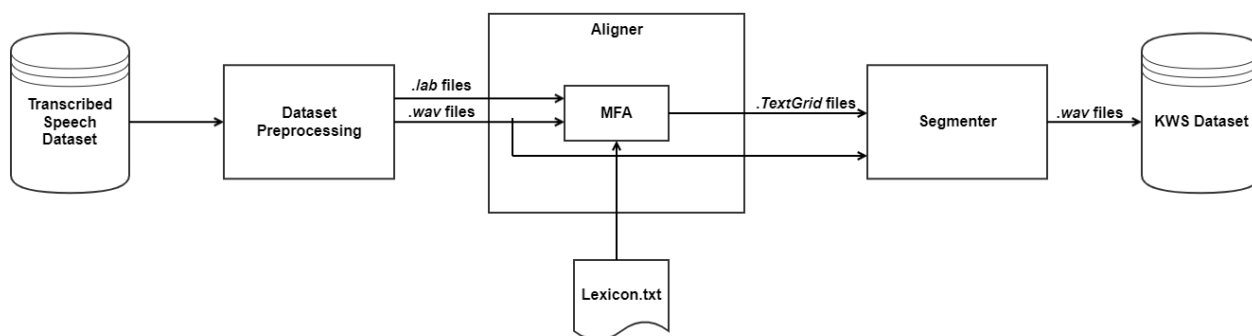


Figure 1: Schematic diagram of KeywordMiner framework pipeline, used to generate SiDi KWS.

The remaining of this document is organized as follows: section 2 introduces the concept of forced alignment, describes the framework Keyword Miner and details the use of that framework to generate the dataset SiDi KWS; section 3 presents some statistics about that dataset and analyses it. Finally, section 4 summarizes how the KWS research community can benefit from the contributions presented by this work.

2. Methodology

2.1. Forced Alignment

2.1.1. Background

Forced Alignment refers to the task of aligning a speech recording with its phonetic transcription [35]. Methods for forced alignment usually rely on automatic speech recognition (ASR) to derive word-level labeling from an audio file given its transcription and a lexical list from its corpus (*i.e.*, a dictionary of pronunciations) [36]. One application of forced alignment is the automatic segmentation of transcribed speech. It has allowed the creation of large training datasets for ASR systems, especially for languages spoken by fewer people [37].

There are popular tools designed to solve that task [38, 39, 40]. Typically, such tools differ on the architectural level and most of them provide pre-trained acoustic models to align utterances based on different languages. Others also provide the choice to train an acoustic model from the ground up with different datasets [34].

2.1.2. MFA

Montreal Forced Aligner (MFA) [34] is an open-source aligner built on top of Kaldi [41], an open-source ASR toolkit. In the background, it implements a Gaussian Mixture Model-Hidden Markov Model to identify utterances. As the audio feature representation, it uses Mel-Frequency Cepstral Coefficients (MFCC). MFA supports the alignment of audio datasets either by training an acoustic model from scratch or by using pre-trained acoustic models attached to its online documentation. Along with the input audio files, MFA must receive their respective transcriptions as text files with the *lab* extension as well as the phoneme dictionary (*i.e.*, lexicon) from the language to be aligned. The tool outputs the aligned utterances in text files with the *TextGrid* extension.

2.2. Keyword Miner

To automate the application of MFA on public datasets of transcribed audio files, this work implemented KeywordMiner: an open-source¹ framework that outputs audio clips of single-spoken keywords segmented via MFA from input datasets of transcribed speech. Figure 1 illustrates its operation. The next subsections describe its components.

2.2.1. Dataset Preprocessing

This block is composed of the abstract class *Dataset*, which interfaces the framework with the implementation of different input speech datasets. It is responsible for arranging the input transcriptions into *lab* files, which correspond to the transcription of each single input audio file. This same block also converts the audio files to the WAV format and standardizes their sample rate to 16 kHz.

2.2.2. Aligner

The *lab* and WAV files generated by the previous block are fed into the aligner block, which runs MFA in the background. Given those files, along with the desired acoustic model and lexicon file, MFA runs its ASR to align the utterances to their transcripts. Afterward, it exports a *TextGrid* file for each input audio file. Each *TextGrid* file stores the alignment pair (*keyword*, *timestamp*), where the timestamp marks the beginning and the end of its respective keyword.

2.2.3. Segmenter

Finally, given the outputs obtained in the previous steps, this block segments each keyword within the WAV files based on their respective *TextGrid* files. Those segmented keywords are exported as audio clips of single-spoken keywords, named with their respective label and creation timestamp. KeywordMiner uses Librosa [42] to both segment and export those audio clips.

2.3. SiDi KWS Creation

SiDi KWS was generated by running KeywordMiner based on the following resources:

- **Input transcribed speech datasets:** LibriSpeech [26], Mozilla Common Voice [27] and MLS [33].

¹<https://github.com/michel-meneses/keyword-miner>.

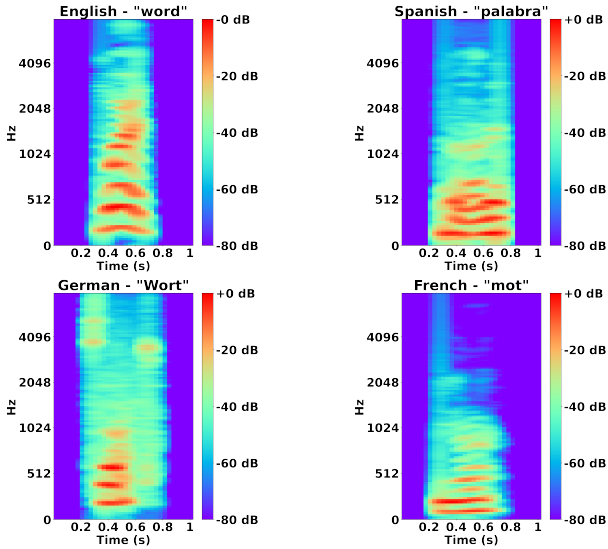


Figure 2: Spectrograms of audio recordings in four languages randomly drawn from SiDi KWS.

- **Input lexicons:** for English and Spanish, this work considered the pre-trained models and lexicons already provided by MFA. As for French and German, new lexicons were generated from scratch by using MFA’s G2P module. Those files worked properly with the French and German acoustic models already provided that aligner.
- **Infrastructure:** this work relied on IARA Lab, the largest supercomputer for artificial intelligence applications in Latin America [43]. For each input speech dataset, KeywordMiner ran based on 128 CPU cores and 1 TB of RAM. It took approximately 720 hours to generate all the labeled recordings that compose SiDi KWS.

3. Results and Discussion

SiDi KWS is publicly available ² in four languages: English, French, German and Spanish. Besides the audio files in the WAV format sampled at 16kHz, it provides metadata containing the audio paths along with their respective keywords, as well as labels indicating whether each audio sample belongs to the suggested train, validation, or test sets. That split follows the ratio 80 : 10 : 10, respectively. This section details the dataset and provides relevant information regarding its main characteristics.

Figure 2 illustrates the time-frequency representation of samples of the keyword “word” spoken in all four languages covered in SiDi KWS. It is important to notice that those particular audio recordings were padded with zeros until completing one second to improve the visualization of those spectrograms. There is no silence either before or after the keywords in the original recordings, as they were extracted using the alignment information, *i.e.*, the positions in which the word begins and ends, provided by MFA.

A dataset designed for training keyword spotters based on deep learning should provide a large number of samples with a high diversity. SiDi KWS is composed of over 24.3 million audio files, from which 791 thousand represent unique keywords. This is a significant improvement in comparison to the currently

²<https://michel-meneses.github.io/sidi-kws/>

public KWS datasets [10, 32, 31]. Table 1 presents the number of unique keywords per language in the dataset, along with the total number of keywords, the ratio between unique and total words, and the average ratio of recordings per keyword. Moreover, Figure 3 presents the distribution of the number of keyword recordings for each language. The language with the largest number of keywords is German, with 10.05 million samples. In terms of variety, French has prevailed, with over 285 thousand unique keywords.

A special case of KWS is the few-shot keyword spotting, in which the learning should be carried out using only a small number of training examples [44, 45]. One way to train few-shot learners is to organize the dataset into tasks, each task emulating the problem that needs to be solved [32]. Thus, if the goal is to solve a few-shot KWS problem using K available examples, each task must contain K training examples of a given word. In light of the growing interest in few-shot solutions, Table 2 presents the number of keywords in SiDi KWS that contain at least five recordings, five being a typical value in few-shot problems.

By comparing all languages against each other based on Table 1 and Table 2, it comes to light that English is the one that lacks variety the most: the percentage of unique keywords is 2.14%, where 46.37% of all English words repeat at least 5 times. This could be a problem for applications that prioritize a large variety of keywords regardless of how much they repeat, although it would be an advantage for applications that need different samples of the same words (*e.g.*, Meta-Learning [46]). A similar distribution can be observed for German, where 2.16% of all the keywords are unique, although the percentage of words repeating at least 5 times is the smallest among all four languages, 22.45%.

The length of the keywords, measured in characters, is distributed across the dataset as indicated in Figure 4 and the average values are presented in Table 3. The available dataset contains all the words extracted by MFA. It is led to the user to decide on filtering the audio samples by the length of their respective keywords or not. German presents most of the longest words, with an average of 8.00 characters per word and a standard deviation of 3.45.

Furthermore, SiDi KWS was also analyzed in terms of the length (in seconds) of its audio files, as shown in Table 4. In essence, the average length of the audio files falls around 0.50s, being English the language with the smaller average and standard deviation: 0.47s and 0.14s, respectively; while German presents the largest, 0.55s and 0.23s. For the whole dataset, the

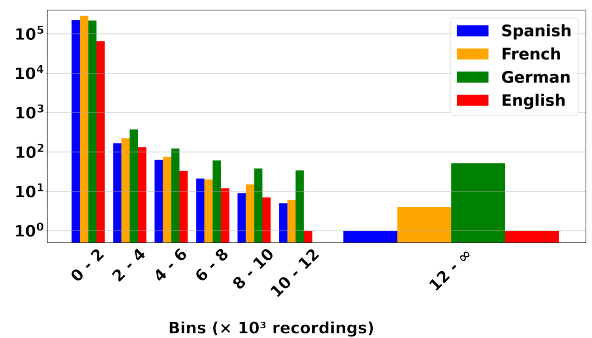


Figure 3: Histogram of recordings per keyword.

Table 1: *Count of audio recordings in SiDi KWS.*

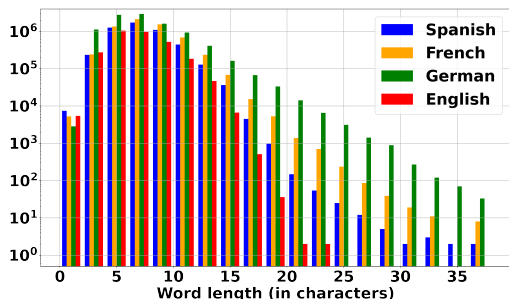
	English	French	German	Spanish	All Languages
Total recordings	3,051,562	6,298,468	10,045,974	4,933,249	24,329,253
Unique keywords	65,199	285,178	217,484	223,333	791,194
Unique / total	2.14%	4.53%	2.16%	4.53%	3.25%
Recordings per word (avg.)	46.8	22.1	46.19	22.1	30.8

Table 2: *Keywords with at least 5 recordings.*

	Count	Ratio
English	30,234	46.37%
French	67,370	23.62%
German	68,789	22.45%
Spanish	58,242	26.08%
All languages	224,635	28.39%

Table 4: *Duration (in seconds) of the audio recordings.*

	Mean \pm Std (s)
English	0.475 \pm 0.137
French	0.513 \pm 0.169
German	0.553 \pm 0.227
Spanish	0.513 \pm 0.179
All languages	0.525 \pm 0.196

Figure 4: *Histogram of keyword lengths.*

average length and standard deviation computed were 0.52s and 0.20s, respectively.

4. Conclusion

This work introduced SiDi KWS, a public large-scale multilingual dataset for keyword spotting, currently composed of 24.3 million labeled audio recordings of single-spoken keywords. It also presented KeywordMiner, an open-source framework specially designed to generate SiDi KWS by applying forced alignment on public datasets of transcribed speech. Both SiDi KWS and KeywordMiner are publicly shared to benefit the speech processing research community. For future work, the authors intend to increment SiDi KWS by running KeywordMiner on new public datasets of transcribed speech in different languages. It is also desired to use SiDi KWS to assess the performance of

Table 3: *Length (in characters) of the keywords.*

	Mean \pm Std
English	6.12 \pm 2.22
French	7.52 \pm 2.63
German	8.00 \pm 3.45
Spanish	7.23 \pm 2.54
All languages	7.20 \pm 2.90

state-of-the-art keyword spotters.

5. Acknowledgements

The results presented in this paper have been achieved as part of a project executed by SiDi Institute and financed by Samsung Eletrônica da Amazonia Ltda., under the auspices of the Brazilian Federal Law of Informatics n°. 8248/91.

6. References

- [1] I. Szöke, P. Schwarz, P. Matejka, L. Burget, M. Karafiát, M. Fapso, and J. Cernocký, “Comparison of keyword spotting approaches for informal continuous speech.” in *INTERSPEECH*, 2005, pp. 633–636.
- [2] J. Tejedor, D. Toledano, P. Lopez-Otero, L. Docio-Fernandez, M. Peñagarikano, L. Rodriguez-Fuentes, and A. Moreno-Sandoval, “Search on speech from spoken queries: the multi-domain international albayzin 2018 query-by-example spoken term detection evaluation.” *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, pp. 1–29, 2019.
- [3] Y. Chen, S. Huang, C. Shen, H. Lee, and L. Lee, “Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval,” in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 941–948.
- [4] A. Shrivastava, A. Kundu, C. Dhir, D. Naik, and O. Tuzel, “Optimize what matters: Training dnn-hmm keyword spotting model using end metric,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4000–4004.
- [5] H. Liu, A. Abhyankar, Y. Mishchenko, T. Sénéchal, G. Fu, B. Kulis, N. Stein, A. Shah, and S. Vitaladevuni, “Metadata-aware end-to-end keyword spotting,” in *INTERSPEECH*, 2020, pp. 2282–2286.
- [6] C. Yang, X. Wen, and L. Song, “Multi-scale convolution for robust keyword spotting,” in *INTERSPEECH*, 2020, pp. 2577–2581.
- [7] T. Higuchi, A. Gupta, and C. Dhir, “Multi-task learning with cross attention for keyword spotting,” *arXiv preprint arXiv:2107.07634*, 2021.
- [8] J. Lin, K. Kilgour, D. Roblek, and M. Sharifi, “Training keyword spotters with limited and synthesized speech data,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7474–7478.
- [9] T. Higuchi, S. Saxena, M. Souden, T. Tran, M. Delfarah, and C. Dhir, “Dynamic curriculum learning via data parameters for

- noise robust keyword spotting,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6848–6852.
- [10] P. Warden, “Speech commands: A dataset for limited-vocabulary speech recognition,” *arXiv preprint arXiv:1804.03209*, 2018.
- [11] Y. Zhang, N. Suda, L. Lai, and V. Chandra, “Hello edge: Keyword spotting on microcontrollers,” *arXiv preprint arXiv:1711.07128*, 2017.
- [12] G. Chen, C. Parada, and G. Heigold, “Small-footprint keyword spotting using deep neural networks,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4087–4091.
- [13] T. N. Sainath and C. Parada, “Convolutional neural networks for small-footprint keyword spotting,” in *INTERSPEECH*, 2015.
- [14] A. Michaely, X. Zhang, G. Simko, C. Parada, and P. Aleksic, “Keyword spotting for google assistant using contextual speech recognition,” in *Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 272–278.
- [15] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [16] A. Chattopadhyay, P. Hassanzadeh, and S. Pasha, “Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data,” *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.
- [17] P. Tirado-Martin and R. Sanchez-Reillo, “Bioecg: Improving ecg biometrics with deep learning and enhanced datasets,” *Applied Sciences*, vol. 11, no. 13, p. 5880, 2021.
- [18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [19] J. Ahmad, H. Farman, and Z. Jan, “Deep learning methods and applications,” in *Deep learning: convergence to big data analytics*. Springer, 2019, pp. 31–42.
- [20] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting unreasonable effectiveness of data in deep learning era,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.
- [21] A. Althnani, D. AlSaeed, H. Al-Baity, A. Samha, A. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, “Impact of dataset size on classification performance: an empirical evaluation in the medical domain,” *Applied Sciences*, vol. 11, no. 2, p. 796, 2021.
- [22] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.
- [23] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and L. Zitnick, “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.
- [24] A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [25] N. Asghar, “Yelp dataset challenge: Review rating prediction,” *arXiv preprint arXiv:1605.05362*, 2016.
- [26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [27] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.
- [28] H. Shin, H. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. Summers, “Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning,” *Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.
- [29] S. Choi, S. Seo, B. Shin, H. Byun, M. Kersner, B. Kim, D. Kim, and S. Ha, “Temporal convolution for real-time keyword spotting on mobile devices,” *arXiv preprint arXiv:1904.03814*, 2019.
- [30] C. Shan, J. Zhang, Y. Wang, and L. Xie, “Attention-based end-to-end models for small-footprint keyword spotting,” *arXiv preprint arXiv:1803.10916*, 2018.
- [31] Ł. Lepak, K. Radzikowski, R. Nowak, and K. Piczak, “Generalisation gap of keyword spotters in a cross-speaker low-resource scenario,” *Sensors*, vol. 21, no. 24, p. 8313, 2021.
- [32] J. Wang, Y. He, C. Zhao, Q. Shao, W. Tu, T. Ko, H. Lee, and L. Xie, “Auto-kws 2021 challenge: Task, datasets, and baselines,” *arXiv preprint arXiv:2104.00513*, 2021.
- [33] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, “Mls: A large-scale multilingual dataset for speech research,” *arXiv preprint arXiv:2012.03411*, 2020.
- [34] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldii,” in *INTERSPEECH*, vol. 2017, 2017, pp. 498–502.
- [35] C. Batista, A. Dias, and N. Neto, “Free resources for forced phonetic alignment in brazilian portuguese based on kaldii toolkit,” *EURASIP Journal on Advances in Signal Processing*, vol. 2022, no. 1, pp. 1–32, 2022.
- [36] C. DiCanio, H. Nam, D. Whalen, H. Timothy, J. Amith, and R. García, “Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment,” *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2235–2246, 2013.
- [37] J. Leinonen, S. Virpioja, and M. Kurimo, “Grapheme-based cross-language forced alignment: Results with uralic languages,” in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2021, pp. 345–350.
- [38] K. Gorman, J. Howell, and M. Wagner, “Prosodylab-aligner: A tool for forced alignment of laboratory speech,” *Canadian Acoustics*, vol. 39, p. 192, 2011.
- [39] R. Fromont and J. Hay, “Labcat: An annotation store,” in *Australasian Language Technology Association Workshop*, 2012, pp. 113–117.
- [40] J. Goldman, “Easyalign: an automatic phonetic alignment tool under praat,” in *INTERSPEECH*, 2011.
- [41] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., “The kaldii speech recognition toolkit,” in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, no. CONF. IEEE Signal Processing Society, 2011.
- [42] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.
- [43] SiDi, Mar 2022. [Online]. Available: <https://www.sidi.org.br/en/sidi-apresenta-o-iara-maior-supercomputador-de-ia-do-pais/>
- [44] M. Mazumder, C. Banbury, J. Meyer, P. Warden, and V. Reddi, “Few-shot keyword spotting in any language,” *arXiv preprint arXiv:2104.01454*, 2021.
- [45] Y. Chen, T. Ko, L. Shang, X. Chen, X. Jiang, and Q. Li, “An investigation of few-shot learning in spoken term classification,” *arXiv preprint arXiv:1812.10233*, 2018.
- [46] A. Parnami and M. Lee, “Few-shot keyword spotting with prototypical networks,” *arXiv preprint arXiv:2007.14463*, 2020.