Artificial Intelligence Research and Development
T. Alsinet et al. (Eds.)

doi:10.3233/FAIA240421

Automatic Catalan Keyword Spotting Database Generator

Àlex SÁNCHEZ [a], Sergio MONTOYA [b], Josep ESCRIG [c], Ivan HUERTA [d]
[a] *alex.sanchez@i2cat.net*
[b] *sergio.montoya@i2cat.net*
[c] *josep.escrig@i2cat.net*
[d] *ivan.huerta@i2cat.net*

Abstract. Home assistants are essential today, but they typically support only popular languages. Promoting products that enhance underrepresented languages is crucial for preservation. Using a home assistant in one's native language, such as Catalan, is a significant step toward this goal. Keyword spotting (KWS) and speech recognition are two potential solutions. The lightweight architecture of KWS models is promising for low-powered edge devices in domotic environments. However, there is a lack of resources to train such models, especially for Catalan. This paper presents a solution using forced alignment techniques with speech-to-text models to extract any set of words from any speech resource. While our focus is on Catalan, this methodology can be applied to other languages.

Keywords. keyword spotting, speech recognition, forced alignment, automatic keyword extraction, edge home assistants

Introduction

In speech processing, command understanding involves recognizing spoken instructions, from simple words like "start" to phrases like "Raise the blinds." With the rise of voice assistants like Amazon's Alexa, command understanding technologies are widespread. Two primary AI approaches address this: Automatic Speech Recognition (ASR) translates spoken sentences into text, while Keyword Spotting (KWS) identifies specific target words in audio. ASR is complex and generally cloud-based, while KWS is less demanding and suitable for edge devices [1]. Large datasets are needed for models to understand commands in specific languages, excluding underrepresented ones like Catalan. Although some Catalan speech databases exist, single-word databases for KWS are rare. Few have emerged, such as Mozilla Common Voice's single-word split [2], with 14 words and 3666 samples, and MLCommons automatic extractions in the MSWC [3], with 29741 words and 31041 samples. We aim to expand this by gathering major Catalan speech resources and extracting specific words to train a home assistant. Our methodology uses transcription techniques to extract words from any speech resource.

In a cloud-connected world, edge domotic systems [1] are valued for their privacy features. Recently, embedded GPU-powered systems have enhanced KWS capabilities [4], but GPUs aren't always available. Running models on simpler processors requires extensive and varied databases [5], which may not exist for underrepresented languages. MLCommons used forced alignment techniques to create a keyword database from speech transcriptions in their Multilingual Spoken Words Corpus [3], but this resulted in a small, varied database. Forced alignment, which aligns phonetic transcriptions with audio recordings, can be done using Montreal Forced Alignment (MFA) [6]. MFA requires a pronunciation dictionary and an ASR model. There are Catalan pronunciation dictionaries and some ASR models, but existing models aren't ideal for alignments. We address the lack of transcribed speech resources by extracting words from non-transcribed ones. We validate our approach with the false alarm rate benchmark [7] to evaluate its robustness against false positives.

Our final database includes 29 words in the domotic domain totalling 405057 samples, making it more suitable for training than the previously existing ones. Our contributions include:

- Training a lightweight ASR model (Hidden-Markov Model [8]) to use Montreal Forced Alignment [6] for speech-transcription alignments.
- Generating an automatic pipeline to create a single-word database from any speech resource.
- Demonstrating the feasibility of automatically extracting a Catalan single-word database and training a KWS model for low-resource edge devices. We address domain-shift challenges with an audio filter that emulates these devices frequency response.

Methodology

The first step is to gather speech resources and their transcriptions. An Automatic Speech Recognition (ASR) model and a pronunciation dictionary are then obtained. In most languages, the ASR can be directly obtained from online resources. In Catalan, contrarily, we trained our own from scratch. Then, the speech data is aligned with the transcriptions using the ASR model and dictionary. Following alignment, a single-word database is created, and a KWS model is trained with these words. This forms the backbone of the pipeline for speech resource utilization and KWS model development. To address the availability of transcriptions, we implemented two pipelines. The basic pipeline starts with transcribed speech resources and applies the specified methodology (see Figure 1). We added a transcription step using OpenAI's Whisper [9] for resources without transcriptions, highlighted in red in Figure 1. Four transcribed resources are used: MCV [2] (2400 hours (h), 34628 speakers (s), ParlamentParla (211h, 379s), TV3Parla (280h, no speaker diarisation), and SLR69 (9.4h, 36s). This data trained the ASR model with MFA's utility. Diverse voices improve ASR results, so speaker labeling is beneficial. For untranscribed resources, we used open-licensed speech resources, including audiobooks , radio shows , and political speeches

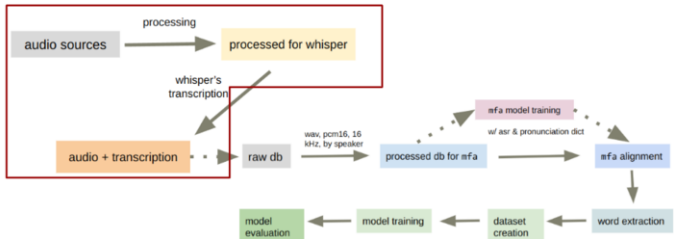

Figure 1. Pipeline: automatic word extraction from speech resources.

After alignment, the initial step in constructing the single-word database for a home assistant involves selecting key words categorized into actions, complements, and levels, structured as *action* + *complement* + *level* (if applicable). Actions: *activa, ajuda, apaga, encén, para, posa, obre, puja, baixa*; levels: *zero, un, dos, tres, quatre, cinc, sis, set, vuit, nou, deu, vint-i-cinc, cinquanta, setanta-cinc, cent*; and complements: *calor, cuina, fred, graus, habitació, jardí, llum, menjador, totes, tots, sala*

Edge Catalan keyword-spotting. The model for this embedded approach needed to be highly accurate yet low-resource demanding. The Multi-head Attention RNN [10], inspired by the "Attention RNN" [11], integrates elements from a CNN, a GRU for better convergence and efficiency [1], and a multi-head attention layer to focus on multiple input segments concurrently. Given the low variety of data available, it is important to significantly augment data synthetically. Data augmentation introduces variability to enhance performance, using techniques like reverb, distortion, pitch shifting, chorus, resampling, acceleration, volume shifting, background volume, and time shifting. Each technique had a 30% probability of being applied during training. To enhance the model's performance in real-world scenarios with limited acoustic quality, we implemented a filter that emulates the frequency response of low-resource microphones. This adjusts the audio signal to match the characteristics of such devices. These data enhancement techniques collectively improve the model's ability to detect words in diverse scenarios. Additionally, we significantly mitigated false positives through negative hard mining [12], which identifies and prioritizes challenging negative samples.

Experimental Results

We experiment training the RNN with the *full* dataset and the *reduced* version. The ACDB includes 11.91 hours of words (42840 samples); 8.65 hours -31140 samples- in the *reduced* version) The test split metrics of the ACDB were promising, but the self-recorded test showed poorer results, indicating models trained on automatically extracted databases may struggle in real-world scenarios due to domain-specific features. However, the *reduced* model showed improved performance despite domain shifts. The False Positive (FP) Benchmark demonstrated the effectiveness of negative hard-mining techniques. More data should increase the model's robustness. With 80% confidence, the test split of the ACDB showed a 92.47% True Positive Rate (TPR) for the *full* version and 98.9% for the *reduced*

one, with 5 and 19 FP, respectively. The *self-recorded low device dataset* showed 57.7% and 78.13% TPR and 124 and 38 FP. The FP Benchmark [7] applied to the *full* version yielded 5 FP/h without negative hard mining, and 2 with it. The *reduced* version went from 2.5 FP/h to less than 1 after applying this technique.

Conclusions

Companies generally avoid investing in products with low returns, limiting language availability. This paper proposes an automatic pipeline for obtaining keyword samples from any speech resource, simplifying the creation of single-word databases for minority languages. We developed a Catalan keyword database, ACDB, achieving over 90% balanced accuracy in KWS. Additionally, we trained a lightweight speech recognition model for low-resource edge devices, successfully tested with low-end microphones and domain-adaptation filtering techniques. Keyword imbalance poses a challenge that can be addressed by extracting more data from more speech resources.

Acknowledgments

This work was supported by the Government of Catalonia

References

[1] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. Hello edge: Keyword spotting on microcontrollers, 2018.
[2] Mozilla common voice. https://commonvoice.mozilla.org/ca/datasets, 2024.
[3] Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, Greg Diamos, Pete Warden, Josh Meyer, and Vijay Janapa Reddi. Multilingual spoken words corpus. In *Thirty-fifth Conference on NeurIPS Datasets and Benchmarks Track*, 2021.
[4] Md Naim Miah and Guoping Wang. Keyword spotting with deep neural network on edge devices. In *2022 IEEE 12th ICEIEC*, pages 98–102, 2022.
[5] Guoguo Chen. Low resource high accuracy keyword spotting, 2016.
[6] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. 2017.
[7] Sergi Sánchez, Ivan Huerta, and Josep Escrig. *A Real-World Dataset for Benchmarking False Alarm Rate in Keyword Spotting.* 10 2023.
[8] Valery Petrushin. Hidden markov models: Fundamentals and applications part 1: Markov chains and mixture models, 2000.
[9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th ICML*, volume 202, pages 28492–28518. PMLR, 23–29 Jul 2023.
[10] Oleg Rybakov, Natasha Kononenko, Niranjan Subrahmanya, Mirkó Visontai, and Stella Laurenzo. Streaming keyword spotting on mobile devices. In *Interspeech*, oct 2020.
[11] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
[12] Jingyong Hou, Yangyang Shi, Mari Ostendorf, Mei-Yuh Hwang, and Lei Xie. Mining effective negative training samples for keyword spotting. In *ICASSP*, 2020.