Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230695

A Real-World Dataset for Benchmarking False Alarm Rate in Keyword Spotting

Sergi SÁNCHEZ DEUTSCH, Ivan HUERTA CASADO and Josep ESCRIG ESCRIG
i2CAT Foundation

Abstract. Over the past few years, Keyword Spotting (KWS) has emerged as a popular area of research. Although numerous open-source KWS datasets have been recently released, there is a general lack of realism in benchmarking the false alarm rate (FAR) in real environments. This can produce models that achieve great accuracies but are not able to work on real-world conditions due to a high number of false triggers. In this work, we demonstrate that two recent KWS models report state-of-the-art accuracies on Google Speech Command dataset but suffer from high false alarm rates in presence of noisy environments. To this end, we propose an extensive benchmark dataset comprising various real-world noises and sounds to evaluate specifically the FAR across different acoustic environments.

Keywords. keyword spotting, command understanding, benchmark, false alarm mitigation

1. Introduction

Keyword Spotting (KWS) is widely used nowadays for wake-up word detection and command recognition. One of the main challenges of KWS systems is to achieve a good detection accuracy of the target keywords while maintaining a low false alarm rate (FAR). An occurrence of a false alarm (FA) takes place when the model identifies a word that was not actually spoken, leading to a wrong action. It is crucial to reduce the number of false alarms to the minimum, as the impact of triggering the device when the keyword is not present is equal or even greater than the cost of not detecting it [1].

In recent years, several datasets have been released for training and benchmarking KWS models. The most well-known is Google Speech Commands (GSC) [2]. This corpus has 10 target keywords along with another 25 words that are used as negative samples to train the *unknown* class. Other less common datasets are Mobvoi [3], which has two target keywords and an extensive set of negative samples composed of other words and speech, and Multilingual Spoken Corpus [4], with 340,000 keywords in 50 different languages.

However, none of the abovementioned datasets include a set of negative samples that is representative enough of the background noises and sounds that can be found in a real environment. This can cause a model to report great accuracies on these datasets while they fail to work in a real environment due to a high number of false alarms per hour.

In this paper, we demonstrate that two state-of-the-art KWS models suffer from high false alarm rates on real acoustic conditions even though they report high accuracies on Google Speech Commands dataset. To this end, we propose an extensive dataset to benchmark specifically the false alarm rate in real environments of any KWS model. This corpus contains more than 52,000 negative samples recorded in a variety of real environments that represent many acoustic situations, and can be used to evaluate FAR of any KWS model.

2. Methods

2.1. FAR Benchmark Dataset

Many open-source KWS benchmarks focus mainly on the ability to discern between the target keywords and other non-target words. However, there is a lack of realism in benchmarking FAR in presence of real-world noises and sounds, so that models are robust enough to work under real conditions. To alleviate this issue, we create an extensive dataset composed of 52,198 1-second audio samples to benchmark specifically the false alarm rate of any KWS model in real-world conditions. These samples contain sounds, speech, and background noises recorded in different real situations. All samples that comprise speech have been manually reviewed to ensure that they do not contain any of the target keywords of the Google Speech Commands dataset. The samples are recorded using a standard laptop microphone at 16 kHz, and are divided in multiple categories, as listed in Table 1: in the category *TV show*, we include samples with several TV contents playing in the background. To gather these samples, the microphone was placed in several locations around the room, and the TV volume was varied several times. The *podcast* category contains a loudspeaker playing several chapters of a BBC podcast. Similarly, the *street* category has a loudspeaker playing noises recorded from a real street in New York City. In the *people* category, we include samples with different crowds talking in the background and, finally, the *living room* category includes multiple daily sounds that can be found in a house, such as people eating at a table or the sound of a vacuum cleaner.

Table 1. Categories and number of samples in our proposed FAR Benchmark dataset

	TV Show	Podcast	Street	People	Living Room
#Samples	23,759	4,713	6,060	3,174	14,492
#Hours	6.6	1.3	1.68	4.03	4.03

2.2. Neural Network Architectures

We conduct the experiments using two state-of-the-art KWS architectures: the Multi-head attention RNN (MHAtt RNN) [5] and the Keyword Transformer (KWT) presented in [6]. The Multi-head attention RNN is based on the “Attention RNN” architecture [7], which combines a CNN, a bidirectional LSTM, and an attention layer. The model has the ability to capture local relations in a feature map and the ability of RNNs to focus on long-term dependencies. The MHAtt RNN model takes the same architecture as the Attention RNN with two main changes, which are a) replacing the LSTM with Gated Recurrent Units (GRU) and b) using the multi-head attention method presented in [8] to focus on more than one relevant part of the input audio. The Keyword Transformer is the first Transformer-based architecture for keyword spotting that relies solely on attention layers. It takes as input the Mel spectrogram of the audio divided into several patches within the time domain. These patches are flattened and mapped to a higher dimension d using a linear projection matrix. Two learnable embeddings are then added to the input patches to provide the encoder layers with some extra information: a class embedding and a positional embedding. These updated inputs (the original spectrogram patches with the extra embeddings) are then fed into a set of L encoder layers, which perform multi-head attention with k different heads. We employ the code released by [9] to train the MHAtt RNN model. For the KWT, we use the implementation released by [6].

3. Experiments

3.1. Dataset

We train the two selected KWS architectures on the Google Speech Commands (GSC) dataset [2]. GSC is a large-scale open-source dataset that comprises over 100,000 utterances of 35 different words recorded by thousands of speakers. It is widely used in research and industry, and it has been key in advancing the field of speech recognition in recent years. From the 35 available words, 10 are used as target keywords and the rest of them are used to train the *unknown* class, as described in Table 2. In our experiments, we use the split files provided by [2] to divide the dataset into training, validation and testing set, with ratio 80:10:10. The unknown class is composed of a subset of 3,500 samples from the non-target utterances to maintain a balance among classes.

Table 2. Target and non-target keywords included in Google Speech Commands dataset.

Target keywords	yes	no	up	down	left
	right	on	off	stop	go
Non-target keywords	zero	one	two	three	four
	five	six	seven	eight	nine
	bed	bird	cat	dog	happy
	house	Marvin	Sheila	tree	wow
	backward	forward	follow	learn	visual

3.2. Results

In Table 3 we evaluate MHAtt RNN and KWT in terms of accuracy on GSC and FAR on our proposed FAR benchmark dataset. The two architectures report state-of-the-art accuracies on GSC, achieving 99.34% in the case of the MHAtt RNN and 99.72% in the case of KWT.

However, FAR results on our proposed dataset show that both models suffer from a serious false alarm problem when facing noisy samples recorded in multiple real environments. If we consider the case of a commercial voice assistant, it would be unacceptable that the system gets triggered by a non-target utterance more than 100 times every hour in the best case scenario (considering the MHAtt RNN). In the case of KWT, this

FAR increases up to 902 FA per hour. Overall, this demonstrates that the robustness of a KWS model can not be benchmarked with enough confidence using only the accuracy on GSC, as it may lead to a FA issue in real-life conditions.

Table 3. Accuracy on GSC and FAR results in our proposed FAR Benchmark dataset.

Model	Accuracy (GSC)	FAR (%)	FAR (FA/h)
MHAtt RNN	99.34	2.88	103.6
KWT	99.72	25.06	902.1

4. Conclusions and Future Work

In this paper, we demonstrated that current KWS benchmarks are not sufficient to evaluate the robustness of the models in terms of FAR in real-life conditions. To this end, we created an extensive dataset of real-world noises and sounds to benchmark FAR in a variety of acoustic environments. Using this dataset, we proved that two state-of-the-art models suffer from a serious false alarm issue when trained on Google Speech Commands dataset. In future work, we plan to further investigate this issue and propose a false alarm mitigation solution at the dataset level. Our next steps will be thus devoted to reduce FAR in real-world conditions with minor impact on the accuracy of the models.

Acknowledgements

This work has been partially funded by the RED.ES program under the 2021 Grant Program for the Funding of R&D Projects on Artificial Intelligence and other Digital Technologies, with grant number 2021/C005/00152763.

References

[1] Lopez-Espejo I, Tan ZH, Hansen J, Jensen J. Deep Spoken Keyword Spotting: An Overview. IEEE Access. 2021 nov:1-1.

[2] Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. arXiv; 2018.

[3] Hou J, Shi Y, Ostendorf M, Hwang MY, Xie L. Region Proposal Network Based Small-Footprint Keyword Spotting. IEEE Signal Processing Letters. 2019 8;26:1471-5.

[4] Mazumder M, Chitlangia S, Banbury C, Kang Y, Ciro J, Achorn K, et al. Multilingual Spoken Words Corpus. In: Vanschoren J, Yeung S, editors. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1. Curran; 2021. .

[5] Rybakov O, Kononenko N, Subrahmanya N, Visontai M, Laurenzo S. Streaming keyword spotting on mobile devices. INTERSPEECH. 2020 may;2020-Octob:2277-81.

[6] Berg A, O'Connor M, Cruz MT. Keyword Transformer: A Self-Attention Model for Keyword Spotting. In: Proc. Interspeech 2021; 2021. p. 4249-53.

[7] Coimbra de Andrade D, Leo S, Loesener Da Silva Viana M, Bernkopf C. A neural attention model for speech command recognition. ArXiv e-prints. 2018 Aug.

[8] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 2017-Decem. Neural information processing systems foundation; 2017. p. 5999-6009.

[9] Google Research. Streaming Aware neural network models [Source Code]; 2020. Available from: https://github.com/google-research/google-research/tree/master/kws_streaming.