Albayzin Evaluation 2024: Wake-up Word Detection Challenge by Telefónica

Fernando López[1,2] Jordi Luque[2]

[1]Universidad Autónoma de Madrid
[2]Telefónica Innovación Digital

{fernando.lopez, jordi.luque}@telefonica.com

May 21, 2024

Abstract

The Wake-up Word Detection Challenge aims to assess the performance of State-of-The-Art Keyword Spotting systems in addressing various industrial needs such as accuracy, inference delay, and computational load. The challenge dataset includes speech samples with and without the "Okey Aura" WuW keyphrase, presenting challenging conditions including phonetically similar non-WuW samples. Moreover, the dataset provides information regarding speakers' gender, age, accent, and acoustic conditions, allowing for the consideration of biases during training.

The evaluation is organized by Telefónica Innovación Digital and it will be conducted as part of the Iberspeech 2024 conference[1] to be held in Aveiro, Portugal, from 11 to 13 November 2024.

1 Introduction

The popularity of voice-based interfaces has grown tremendously, mainly because it enables hands-free communication with a wide variety of devices. These interfaces' success depends on the efficiency of the on-device wake-up word (WuW) detector. It is a mechanism to identify a specific trigger word or phrase to initiate communication between the user and the device. By detecting the trigger word, the device becomes attentive to the user's request, enabling seamless and smooth interaction.

The accuracy and inference delay of the WuW detector determines the overall usability and effectiveness of voice-based interfaces. Moreover, when considering on-device execution, accounting for computational load and energy effi-

[1]https://iberspeech.tech/

ciency becomes imperative. Additionally, during inference, the detector is often exposed to both mismatch and challenging circumstances such as far-field scenarios, background noise presence, variations in device characteristics, acoustic room environments, and speaker diversity. Thus, reaching robustness in keyword spotting (KWS) presents a significant challenge, especially given that the training datasets do not comprehensively represent all the unforeseen conditions encountered by the KWS model during execution.

The *Albayzin 2024: Wake-up Word Detection Challenge* aims to evaluate Keyword Spotting (KWS) systems under diverse acoustic conditions to address these challenges. This task will assess the performance of State-of-The-Art KWS technology used as a wake-up mechanism for voice assistants.

2 Challenge description and Database

Albayzin 2024: Wake-up Word Detection Challenge consist of detecting the presence of the "Okey Aura" keyphrase in an audio sample. For this evaluation, Telefónica Innovación Digital has licensed a dataset that was initially introduced in [2]. It has been further enhanced with strong temporal alignments of utterances [3]. This dataset contains speech with either the WuW, "Okey Aura", or without it, mixed with domestic background noise samples. The audio clips are designed for streaming performance evaluation. Moreover, each audio file within the dataset has detailed metadata that provides additional context and supports analysis.

2.1 Database

The *Okey Aura Wake-up Word Dataset* comprises 1247 utterances (1.4 hours) from 80 speakers. They are stored in mono-channel Waveform Audio File Format (WAV) by using a Pulse-Code Modulation (PCM) encoded with two bytes per sample at a rate of 16 kHz. The metadata is provided within Tab Separated Values (TSV) format in files with the *.tsv* extension.

This database has been specifically designed to address challenging conditions and therefore includes positive WuW, containing the keyphrase class, and non-WuW samples which are phonetically similar to the keyphrase. Both classes contain real domestic background noises at different Signal-to-Noise Ratio (SNR) levels. The associated metadata provides a comprehensive profile for each utterance, encapsulating a variety of speaker attributes such as age, gender and accent; as well as the specific recording conditions and additional characteristics pertinent to the text transcription. This metadata is systematically catalogued in Table 1. Moreover, Table 2 describes the varying degrees of phonetic similarity of utterances to the WuW.

From the database, we separated samples categorized under the *Exact with context* category due to the need for specific alignments to isolate the keyphrase from the rest of the speech within the utterance. Nonetheless, these samples are included in the `other.tsv` file. The remaining data is compiled into the

Table 1: Metadata in the Okey Aura Wake-up Word Dataset.

Metadata	**Values**
Speaker_ID	Anonymized alphanumeric string of 16 characters
Age	20s, 30s, 40s, 50s, 60s...
Gender	Female, Male, Non-binary
Accent	Andalusian, Andean-pacific, Castilian, Non-native...
Microphone_Distance	Close, Two steps away
Room_size	Small (0 - 10 m^2), Medium (10 - 20 m^2)
Prosody	Unknown, Neutral, Annoyed, Friendly
Transcription	Spoken user utterance
Label	WuW or NonWuW
Similarity	Exact, Exact with context, Okey, Aura...
Audio_Length	Length of the audio in seconds
Start_Time	Start of the spoken content
End_Time	End of the spoken content

Table 2: Phonetic similarity levels in the Okey Aura Wake-up Word Dataset.

Similarity level	**Explanation**	**Example**
Exact	Exactly the WuW	"Okey Aura"
Exact in context	WuW plus an utterance	"*Okey Aura, sube el volumen*"
Okey	Word "Okey" in an utterance	"*Okey, ¿qué te parece?*"
Aura	Word "Aura" in an utterance	"*¡Qué Aura más positiva!*"
Like Aura	Phonetically $\sim$ "Aura"	"*Laura*"
Like Okey	Phonetically $\sim$ "Okey"	"*Hockey*"
Like Okey Aura	Phonetically $\sim$ "Okey Aura"	"*Quiero ver el hockey ahora*"

`validated.tsv` file. From the validated dataset we provide three partitions: `train.tsv`, `dev.tsv`, and `test.tsv`. These partitions consist of stratified training, development and test splits. The partitions are carefully created to ensure that no speaker is repeated between them, maintaining an 80-10-10 ratio.

The domestic background noises audio clips are contained in the `noises.tsv`. Twelve audio files, lasting 15 minutes each. They have been recorded in a living room in Madrid downtown.

In addition, we provide long audio files using the `test.tsv` speech samples and combining them with domestic background noises from `noises.tsv` at different SNR levels. These combined audios are contained in the `test-extended.tsv` file. This data simulates more realistic conditions, as the trigger word is arbitrarily placed within the audio.

2.2 Training and Development Data

Participants can use the whole *Okey Aura Wake-up Word Database* for training and development. Additionally, the participants are free to use any other data to train their model. In such cases, the additional data must be fully documented in the systems description paper. The description of the training data must contain at least the number of hours and the origin of the external data used. For public databases, the name of the database must be provided. For private databases, a brief description of the origin of the data must be provided.

2.3 Reference Result

In addition to the training and development splits, we have included a separate test partition, `test.tsv`. This split allows participants to obtain a reference measurement for the WuW detection task against a baseline system. To assess the model's performance in a more realistic scenario, we also provide `test-extended.tsv`, where participants can evaluate the model's performance on long audio files with higher length variability. In this extended split, both negative and positive samples are combined with domestic background noises to promote diversity in the testing data. The background noise is added to the original waveform at a specific SNR level.

The GitHub repository `https://github.com/ferugit/wuw-challenge-2024` contains a baseline system and simple scripts to assess model performance.

2.4 Evaluation Data

Evaluation data will be similar to the data provided in `test-extended.tsv`. Specifically, it will consist of long audio files, of variable length, containing either positive or negative WuW samples, both of which are combined with domestic background noises at varying SNR ranges.

2.5 Data Organization

The *Okey Aura Wake-up Word Database* is organized as follows:

- `okey-aura-v1.1.0/clips` - a folder with the audio files.
- `okey-aura-v1.1.0/metadata` - a folder with the TSV files.
- `okey-aura-v1.1.0/noises` - a folder with the background noises audio files.
- `okey-aura-v1.1.0/noises/train` - a folder with the background noises audio files that can be used to train the models.
- `okey-aura-v1.1.0/noises/dev` - a folder with the background noises audio files that can be used to validate the models.

- `okey-aura-v1.1.0/noises/test` - a folder with the background noises audio files that can be used to evaluate the models.
- `okey-aura-v1.1.0/extended_test` - a folder with the extended test files.
- `okey-aura-v1.1.0/extended_test/clips` - a folder with the extended test audio files.
- `okey-aura-v1.1.0/extended_test/test-extended.tsv` - a TSV file containing the metadata of how has been produced the extended test audio files.
- `README.md` - a description of the database.
- `EULA_2024` - a copy of the signed EULA to access the database.

3 Performance Measurement

The KWS system will be evaluated using a primary metric to rank the submitted systems. All participants are required to provide, for each audio sample, the probability of the presence of the WuW, along with a binary decision indicating the presence of the WuW. These results must be formatted in a TSV file. Additionally, participants may optionally provide detection timestamps within the long audio file, which will be assessed using an alternative metric.

3.1 Primary Metric

The primary performance measure employed is the Detection Cost Function[1], which is defined as follows:

$$DCF(\theta) = C_{\text{Miss}} \times P_{\text{Miss}}(\theta) \times P_{\text{wuw}} + C_{\text{FA}} \times P_{\text{FA}}(\theta) \times (1 - P_{\text{wuw}}) \quad (1)$$

where θ denotes the decision threshold, the error rates $P_{\text{Miss}}(\theta)$ and $P_{\text{FA}}(\theta)$ are functions of the detection threshold. A false alarm, $P_{\text{FA}}(\theta)$, occurs when the system incorrectly identifies non-target audio as the wake-up word. Conversely, a miss $P_{\text{Miss}}(\theta)$ occurs when the system fails to detect the wake-up word when it is in fact spoken. For a given threshold, these error rates are defined as follows:

$$P_{\text{miss}} = \frac{\text{Number of misses}}{\text{Number of wake-up word samples}} \quad (2)$$

$$P_{\text{FA}} = \frac{\text{Number of false alarms}}{\text{Number of non-wake-up word samples}} \quad (3)$$

The the Detection Cost Function (DCF) represents a weighted sum of the false-reject (miss) and false-accept (false-alarm) error probabilities.

Additionally, the values are given for the target prior, P_{wuw}, and the costs, C_{Miss} and C_{FA}, of, respectively, miss and false-accept errors. P_{wuw} is hypothetical, it should not be confused with the ratio of WuW samples to non-wake-up word samples in the evaluation database. The use of a cost function enables the balancing of different types of errors according to their impact on the application.

A secondary metric, the $minDCF$ see 3.2, provides a single scalar evaluation that reflects the system's best possible performance given the optimal decision threshold.

Table 3: Application-dependant parameters

Parameter	**Value**
P_{wuw}	0.1
C_{Miss}	1
C_{FA}	10

3.2 Alternative Metrics

In addition to the primary metric, other alternative metrics may be computed, although they are not considered for the challenge, as the minimum Detection Cost Function ($minDCF$):

$$minDCF = \min_{\theta}[DCF(\theta)] \tag{4}$$

3.2.1 Timestamps

A secondary metric pertains to the system's ability to temporally localize the WuW within the audio length. Further details regarding this metric will be provided in the future.

4 Evaluation Protocol

This challenge is conducted as an open evaluation where the test data is sent to the participants who process the data locally and submit the output of their systems to the organizers for scoring.

4.1 Registration Rules

All teams willing to participate in this evaluation must be registered through the challenge web page `http://catedrartve.unizar.es/albayzin2024.html` before September 2nd, 2024. In case of any difficulty, you can send an e-mail to `lleida@unizar.es`

4.2 Data License Agreement

The *Okey Aura Wake-up Word Dataset* is available to evaluation participants and is subject to the terms of an End-User Licence Agreement (EULA) with Telefónica Innovación Digital. The data can be retrieved from the online repository[2]. The license agreement can be downloaded from the challenge web page (http://catedrartve.unizar.es/albayzin2024.html).

Participants must sign the agreement (digital signatures are valid) and send a copy attached to an email to the following addresses:
`fernando.lopez@telefonica.com` and `jordi.luque@telefonica.com`. A copy signed by a Telefónica Innovación Digital representative will be returned.

4.3 Evaluation rules

4.3.1 Submission procedure

Each participant team must submit at least a primary system, though they can also submit up to three contrastive systems. Every submitted system must be applied to the entire test database. The evaluation ranking will be based on the results of the primary systems; however, the results of the contrastive systems will also be analyzed and presented during the evaluation session at Iberspeech.

All participants must agree to make their submissions (including system output and system descriptions) available for experimental use by the other participants and the organizing team. Participant teams are required to notify and provide the total time needed to run the set of tests for each submitted system, specifying the computational resources used.

Manual intervention is prohibited in generating system outputs; therefore, all developed systems must be fully automatic. Listening to the evaluation data or any other form of human interaction with the evaluation data is not allowed before all results have been submitted. The evaluated systems must rely solely on automatically processing the provided audio signals.

4.3.2 Results Submission Guidelines

The evaluation results must be presented in just one ZIP file. The ZIP file must contain at least one TSV. Files must adhere to the following naming convention:

Each TSV file must be identified as: $< SITE >< SYSID > .tsv$, where:

- $< SITE >$: Refers to the acronym identifying the participant team (e.g., UPM, UPC, UVI).
- $< SYSID >$: An alphanumeric string identifying the submitted system.

For the primary system, the $< SYSID >$ string must begin with `p-`. For contrastive systems, it should begin with `c1-` for contrastive system 1, `c2-` for contrastive system 2, and `c3-` for contrastive system 3. Each TSV file must

[2] https://zenodo.org/records/11082517

be named as $< SYSID > .tsv$. The zip output file must be identified as $< SITE > .zip$.

Every TSV result file must be formatted to contain the following columns separated by tab:

- `Filename`: the audio filename that has been processed.
- `Probability`: the probability of the audio to contain the WuW.
- `Label`: a binary decision of the audio to contain the WuW, 1 for WuW presence and 0 for WuW absence.
- `Start_Time`: (optional) staring time in seconds of the WuW, "Unknown" in the case of WuW absence.
- `End_Time`: (optional) ennding time in seconds of the WuW, "Unknown" in the case of WuW absence.

Each participant team must upload the zip files through the challenge web page(`http://catedrartve.unizar.es/albayzin2024.html`). In case of uploading problems, participants should send an email with the corresponding ZIP result files to `fernando.lopez@telefonica.com`.

4.3.3 System Descriptions

Participants must send, along with the result files, a PDF file with the description of each submitted system. The format of the submitted documents must fulfil the requirements given in the IberSpeech 2024 call for papers. You can use the templates provided for the Iberspeech conference (WORD or LATEX). Please, include in your descriptions all the essential information to allow readers to understand the key aspects of your systems. A full conference paper can be submitted to the IberSpeech Conference as a regular paper for the Albayzin Evaluation special session. Please, take advise of the deadlines in the IberSpeech 2024 web page https://iberspeech.tech/

5 Schedule

- May 20th, 2024: Registration opens and release of the training data.
- June 3rd, 2024: Release of training and development data.
- July 31st, 2024: Registration deadline.
- September 2nd, 2024: Release of the evaluation data.
- October 18th, 2024: Deadline for submission of results and system descriptions.
- October 31tst, 2024: Results distributed to the participants.

- November 12th, 2024: Official results presented publicly and published
- November 12th, 2024: IberSpeech 2024 special session in Aveiro

References

[1] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech Language*, 20(2):230–275, 2006. Odyssey 2004: The speaker and Language Recognition Workshop.

[2] Guillermo Cámbara, Fernando López, David Bonet, Pablo Gómez, Carlos Segura, Mireia Farrús, and Jordi Luque. Tase: Task-aware speech enhancement for wake-up word detection in voice assistants. *Applied Sciences*, 12(4):1974, 2022.

[3] Fernando López and Jordi Luque. Iterative pseudo-forced alignment by acoustic CTC loss for self-supervised ASR domain adaptation . In *Proc. IberSPEECH 2022*, pages 46–50, 2022.