Contents lists available at ScienceDirect

Nuclear Engineering and Technology

journal homepage: www.elsevier.com/locate/net

Original Article

Towards automatic inspection of nuclear fuel elements in spent fuel pools: Audio analysis

Sergio Segovia [a,*], Ángel Ramos [a], David Izard [a], Doroteo T. Toledano [b]

[a] Development Engineer, Department of Industrial Technology, ENUSA Industrias Avanzadas, Santiago Rusiñol, 28040, Madrid, Spain
[b] Higher Polytechnic School, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, 28049, Madrid, Spain

ARTICLE INFO

Keywords:
Spent fuel pools
Nuclear services
Automatic speech recognition
Audio transcription task

ABSTRACT

In this article, we propose and explore a novel step in the digitization of the mapping of the spent fuel pool of nuclear power plants, in which the audio signal from the operator`s microphone is used to obtain the identification codes of those components that are in each of the cells of the pool. In this way, we have not only an acquisition system but also a verification system that can be used in combination with the outcome of the analysis of the video signal. We developed an algorithm that uses at its core one of the latest models of multilingual Automatic Speech Recognition to transcribe audio signal, and with a post-processing of the timed transcriptions we build the identification code of fuel heads and other components. Results show a very high accuracy in audios from real recording of Spanish nuclear facilities, and the methodology proposed is easily extensible to other nuclear facilities in the world.

1. Introduction

The operation of all types of nuclear reactors generates spent nuclear fuel, which must be safely managed once removed from the reactor core. Spent fuel is considered waste in some circumstances and a potential future energy resource in others. Both management options entail a series of stages, which will necessarily include the storage of spent fuel for a period. The safety aspects of storage are the same as those applied to radioactive waste and are outlined in publication GSR Part 5 [1].

Storage options include wet storage, in some form of pool, and dry storage, in a specially constructed facility or casks. Initially this material is stored in the spent fuel pool (SFP) of the nuclear plant to allow for cooling. The fuel elements are placed in racks at the bottom of the pool, strategically located to limit the separation distances between fuel elements to maintain the required temperature and critical conditions. The International Atomic Energy Agency (IAEA) stipulates in its technical guidelines that each reactor should maintain a database providing uniform and standardized data on a common basis and with an appropriate level of detail [2].

These data can also be combined to track and estimate inventories of nuclear materials. The most basic data for spent fuel management includes inventories of spent fuel in terms of quantity, location, and their characteristics, often used for analysis or planning. These data are at least required at the national level by the regulatory body, which in Spain is the Consejo de Seguridad Nuclear (CSN).

In the CSN Safety Guide 1.7 [3] regarding the information to be provided by the operator of the nuclear facility, particularly related to irradiated fuel stored in the pool, it is required to include (among many other details).

- **The identification of the fuel element:** A unique identification code for each element, assigned by the manufacturer. This identification code should be unique at least within the inventory of each individual reactor, and usually consists of two letters and two digits.
- **The location within the pool.**

As a result of these technical instructions by regulatory bodies, it is necessary to carry out a task known as pool mapping, registering the fuel elements and any other components present in each of the cells of the SFP racks.

Currently, pool mapping in Spanish SFPs is performed by a team of operators divided into two groups. The first one uses an underwater video camera attached to a pole, which is moved through each cell of the pool racks using a crane bridge. The operator positions the camera focusing on the interior of the cell to visualize its contents, and particularly their identification codes. The second group of operators is

* Correspondig author.
E-mail addresses: sergio.segoviag@estudiante.uam.es (S. Segovia), rag@enusa.es (Á. Ramos), dih@enusa.es (D. Izard), doroteo.torre@uam.es (D.T. Toledano).

https://doi.org/10.1016/j.net.2024.05.008
Received 22 November 2023; Received in revised form 19 April 2024; Accepted 9 May 2024
Available online 11 May 2024

located at the edge of the pool, where one person is responsible for adjusting the camera's focus and lighting, while recording an audio signal with keywords including coordinates of the cell being viewed and identification codes of the fuel elements and other components observed in the real-time video signal. Other operators in this second group are responsible for transcribing the codes into a file and verifying that the observed component matches what is expected according to the pool map provided by the nuclear facility at the beginning of the campaign. This operation is highly susceptible to human errors. Therefore, additional double checks that do not increase costs substantially are very valuable.

In this article we propose, explore and validate the partial automation of the creation of pool maps in Spanish nuclear facilities, by using state-of-the-art Automatic Speech Recognition (ASR) technology and evaluating its performance in real data obtained in the normal pool mapping operation of all the nuclear plants in Spain. Previous works have already proposed the idea of using ASR for the partial automatization of spent nuclear fuel. For instance, in Ref. [4] the authors present a demo mobile app that could help transcribe speech for different tasks, mentioning spent nuclear fuel inspection as one of the possible applications, but no evaluation on real data is presented. In a recent survey of deep learning usage in the nuclear industry [5], the authors mention the possibility of processing audio for different tasks in the nuclear industry, but do not mention explicitly spent fuel management as one of them, just suggesting that deep learning in general can be applied to this problem.

The main contributions of our article are therefore.

- To the best of our knowledge, this article presents the first complete application and thorough evaluation of state-of-the-art ASR technology in the context of spent nuclear fuel management on real data.
- Results show a high level of precision, robustness and reliability, demonstrating that the current ASR technology can be applied to this problem, either for the partial automation of the spent fuel pool mappings or as a double check to reduce the probability of human errors.
- The system developed can operate in real time and includes ASR transcription and a specific post-processing component that are described in detail.
- Given the architecture of the system proposed and the capabilities of the ASR system employed, we argue that our proposed architecture can be easily applied to other facilities and other languages worldwide.

The rest of this article is organized as follows. Section 2, briefly introduces the current state of the art in Automatic Speech Recognition. Section 3, describes the election of the base model for transcription of audio signal. Results obtained on stored videos are shown in Section 4. A discussion about future improvements and conclusions are presented in Section 5 and Section 6 respectively.

2. State of art in Automatic Speech Recognition

Automatic speech recognition (ASR) models are used for the task of transforming speech to text, often referred to as transcription. The accuracy of these models have been significantly boosted since deep neural network (DNN) based hybrid model [6] was adopted a decade ago. This breakthrough used DNN to replace the traditional Gaussian mixture model for the acoustic likelihood evaluation, while keeping all the components such as the acoustic model, language model, and lexicon model, etc., as a hybrid ASR system. More recently, the speech community had a new breakthrough by transiting from hybrid modeling to end-to-end (E2E) neural models [7,8] which directly translates an input speech sequence into an output token sequence using a single neural network.

Models have achieved human parity [9] in certain tasks, which means that models are able to transcribe human speech with the same or better accuracy than humans. However, due to the large amounts of data required by DNN-based models, most of these advances have been restricted to a few languages. Besides, performance degraded importantly when testing in a different corpus.

Recently, there has been a paradigm shift in Natural Language Processing (NLP) research with the advent of the Transformer model [10]. These models have found application as pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT) [11]. These models first use unlabeled data (raw text) for pre-training and are then fine-tuned for a downstream task using labeled task-specific data. The intuition behind this approach is that the model can learn features from unlabeled data that can be shared across tasks. This idea has been quickly exported to other fields, including ASR.

2.1. Multilingual ASR models

Multilingual ASR tackles the task in multiple languages with the same model or pipeline. Multilingual Language Models [12] are pre-trained on a large amount of unlabeled text from multiple languages (around 100) and then fine-tuned for a particular language and task. These models have shown impressive cross-lingual transfer capabilities, leading to performance gains on languages that have no labeled data (through zero-shot learning) or a small amount of labeled data (few-shot learning).

Recently, self-supervised learning (SSL) has been used to pre-train an encoder on unlabeled data using contrastive loss. In this setup, the model learns a general high-level contextual representation of the input data which can potentially be used for any downstream task. The training process is divided into two steps [13].

- Learning an encoder which maps the raw audio to a high-level compact representation, usually done using Convolutional Neural Networks (CNNs).
- Reconstruct the future frames given the high-level features, using a strong auto-regressive model such as a Transformer.

These pre-trained audio encoders, known as Wav2Vec2.0, learn high-quality representations of speech, but because they are purely unsupervised, they lack an equivalently performant decoder mapping those representations to useable outputs, necessitating a fine-tuning stage to perform a task such as speech recognition. The two most downloaded ASR models [14,15] in the popular AI platform huggingface (huggingface.co) at the time of writing this manuscript were derived from this approach.

2.2. Whisper. Multilingual transcription model

An alternative to fine-tunning purely unsupervised models is the one proposed by OpenAI with their Whisper models [16]. These models, released in 2023, are Transformer models trained on weakly supervised data, but with an order of magnitude more data than previous systems: 680,000 h of labeled audio data, equivalent to over 70 years of continuous speech. This allows these models to work well with existing datasets and new data even with a zero-shot approach, removing the need for any dataset specific fine-tuning to achieve accurate results.

These models are trained for speech recognition and speech translation tasks, and they can be in 5 different sizes, summarized in Table 1. Over many existing ASR systems, Whisper models exhibit improved robustness to accents, background noise, and technical language, as well

Table 1
Comparison of the versions of Whisper. Parameters is the number of trainable parameters in millions.

VERSION	Tiny	Base	Small	Medium	Large
PARAMETERS	39	74	244	769	1550

as zero-shot translation from multiple languages into English, and their accuracy on speech recognition and translation is near the state-of-the-art level even in a zero-shot approach, and also in par with expert human transcribers. In that sense, Whisper has opened a new level of performance and applicability of ASR technology establishing a new state-of-the-art. The most recent developments, such as Google Gemini [17], released in December 2023, compares its performance against the state-of-the-art established by Whisper. One of the Whisper models (Whisper-Large-v3) [18] is the trendiest ASR model, and the 6th most downloaded model in the popular AI platform huggingface (huggingface.co) at the time of writing this manuscript.

However, because the models are trained in a weakly supervised manner using large-scale noisy data, predictions may include text that is not actually spoken in the audio input (e.g., hallucinations). This is thought to occur because, given their general knowledge of language, the models combine the attempt to predict the next word in the audio with the attempt to transcribe the audio itself. It is a common situation in natural language processing with these models, even in Large Language Models (LLMs) such as Chat-GPT3.

Whisper models perform unevenly across languages, and we observe lower accuracy on low-resource and/or low-discoverability languages or languages where there is less training data. They also exhibit disparate performance on different accents and dialects of languages, which may include a higher word error rate across speakers of different genders, races, ages, or other demographic criteria. In addition, the sequence-to-sequence architecture of the model makes it prone to generating repetitive texts, which can be mitigated to some degree by beam search and temperature scheduling.

Another common problem in the ASR systems is that they employ heuristic sliding window style approaches, which are prone to errors due to overlapping or incomplete audio (e.g., words being cut halfway through). Whisper, in this case, proposes a buffered transcription approach that relies on accurate timestamp prediction to determine the amount to shift the subsequent input window by, but such a method is prone to severe drifting since timestamp inaccuracies in one window can accumulate to subsequent windows. The hand-crafted heuristics employed have achieved limited success.

For these reasons new models have been developed, which are considered variants of the Whisper model since they are based on the use of these as a base, and new functions are added to solve or improve as far as possible the problems that these models present, without losing their high accuracy.

2.2.1. *WhisperX*

One of such new models is WhisperX [19], a system for efficient speech transcription of long-form audio with accurate word-level timestamps. It consists of three additional stages to Whisper transcription.

- Pre-segmenting the input audio with an external Voice Activity Detection (VAD) model.
- Cutting and merging the resulting VAD segments into approximately 30 s input chunks with boundaries lying on minimally active speech regions enabling batched whisper transcription.
- Forced alignment with an external phoneme model to provide accurate word-level timestamps.

These approaches have several drawbacks, including.

* The need to find one wav2vec model per language to support, which does not scale well with the multi-lingual capabilities of Whisper.
* The need to handle (at least) one additional neural network (wav2vec model), which consumes memory.
* The need to normalize characters in Whisper transcription to match the character set of the wav2vec model. This involves awkward language-dependent conversions, such as converting numbers to words ("2" → "two"), symbols to words ("%" → "percent", "e" → "euro(s)").
* The lack of robustness around speech disfluencies (fillers, hesitations, repeated words, etc.) that are usually removed by Whisper.

2.2.2. *Whisper-timestamped*

One solution to the drawbacks of the WhisperX model is the approach based on Dynamic Time Warping (DTW) [20] applied to cross-attention weights, which does not have these drawbacks. DTW is the name of a class of algorithms for comparing series of values with each other. The rationale behind DTW is, given two time series, to stretch or compress them locally to make one resemble the other as much as possible. The distance between the two is computed, after stretching, by summing the distances between individually aligned elements. This technique is useful, for example, when one is willing to find a low distance score between the sound signals corresponding to utterances "*now*" and "*nooow*" respectively, insensitive to the prolonged duration of the/o/sound.

One model that employs this approach is Whisper-timestamped [21], which we will use in this article. As in the previous case, it is a model that can be considered as a variant of the Whisper models, because it is an implementation to predict word timestamps and provide a more accurate estimation of speech segments when transcribing with Whisper models.

3. Proposed methods

We have chosen to use the Whisper-timestamped model, described in the last section, using the *'large'* version of the Whisper model (see Table 1) as model base. It obtains the transcription of the audio with great precision, and more useful information for our task, such as the timestamps of key words, which are of interest for our final purpose.

To extract the information of interest from the complete audio transcript (i.e., the cell coordinates and the codes of the fuel elements, or of other components) we make use of keywords that the operator always says during the recording, which indicate the time from which the codes are said. In addition, the use of the International Radiotelephony Spelling Alphabet (also known as NATO phonetic alphabet) facilitates the identification of characters that are letters. Fig. 1 shows the flow diagram of the proposed method, detailing the way in which the developed system, starting from an audio signal, extracts the required information.

3.1. *Word error rate (WER)*

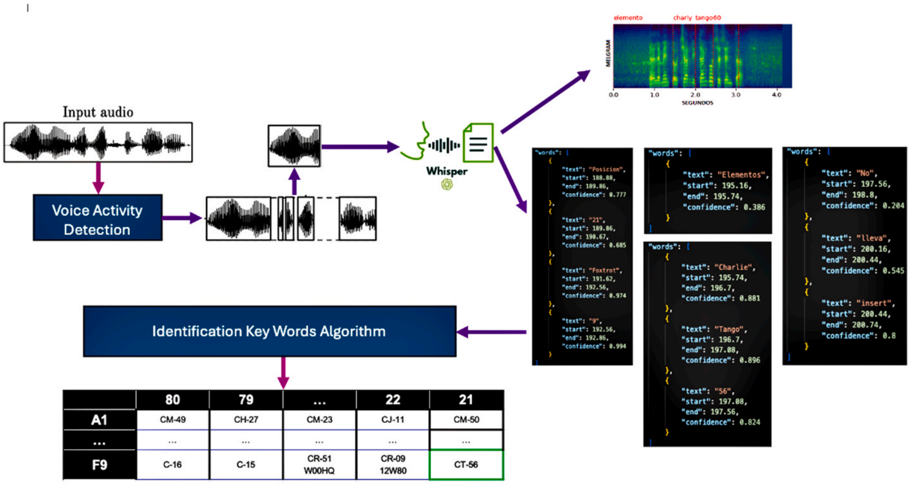
The Word Error Rate (WER) is the most common metric used to evaluate the accuracy of an ASR system, it represents the transcription error rate by comparing the total number of incorrect, deleted, or inserted words with respect to the reference transcription. It is expressed as a percentage and is calculated using the formula:

$$WER = \frac{S + D + I}{N} \times 100 \tag{1}$$

Where S is the number of substitutions (incorrect words), D is the number of deletions (omitted words), It is the number of insertions (inserted words), and N is the total number of words in the reference transcription.

4. Transcription in video files

First, to test the selected model, we used stored videos, not an audio signal in real time, from previous pool mappings in five Spanish nuclear power plants (NPP-1, NPP-2, NPP-3, NPP-4, NPP-5). An example of a transcription obtained by a person (reference) and a transcription obtained by the model (transcription) is shown below.



```
"words": [
    {
        "text": "Posicion",
        "start": 188.88,
        "end": 189.86,
        "confidence": 0.777
    },
    {
        "text": "21",
        "start": 189.86,
        "end": 190.67,
        "confidence": 0.685
    },
    {
        "text": "Foxtrot",
        "start": 191.62,
        "end": 192.56,
        "confidence": 0.974
    },
    {
        "text": "9",
        "start": 192.56,
        "end": 192.86,
        "confidence": 0.994
    }
]
```

```
"words": [
    {
        "text": "Elementos",
        "start": 195.16,
        "end": 195.74,
        "confidence": 0.386
    }
]
```

```
"words": [
    {
        "text": "Charlie",
        "start": 195.74,
        "end": 196.7,
        "confidence": 0.881
    },
    {
        "text": "Tango",
        "start": 196.7,
        "end": 197.08,
        "confidence": 0.896
    },
    {
        "text": "56",
        "start": 197.08,
        "end": 197.56,
        "confidence": 0.824
    }
]
```

```
"words": [
    {
        "text": "No",
        "start": 197.56,
        "end": 198.8,
        "confidence": 0.204
    },
    {
        "text": "lleva",
        "start": 200.16,
        "end": 200.44,
        "confidence": 0.545
    },
    {
        "text": "insert",
        "start": 200.44,
        "end": 200.74,
        "confidence": 0.8
    }
]
```

	80	79	...	22	21
A1	CM-49	CH-27	CM-23	CJ-11	CM-50
...	...	...	...	...	...
F9	C-16	C-15	CR-51 W00HQ	CR-09 12W80	CT-56

Fig. 1. Flowchart of the proposed method.

- **Reference** - "… posicion 30 bravo delta elemento alpha hotel 42 no lleva insert posicion 30 bravo charlie elemento alpha lima 59 no lleva insert posicion 30 bravo bravo elemento sierra alpha 44 no lleva insert posicion 30 alpha tango elemento alpha golf 51 no lleva insert posicion 30 alpha sierra elemento alpha golf 35 no lleva insert posicion 30 alpha romeo elemento alpha golf 17 no lleva insert posicion 30 alpha quebec elemento alpha echo 03 no lleva insert, …"
- **Transcription** – "… posicion 30 bravo delta elemento alfa hotel 42 no lleva insert posicion 30 bravo charly elemento alfa lima 59 no lleva insert posicion 30 bravo bravo elemento sierra alfa 44 no lleva insert posicion 30 alfa tango elemento alfa golf 51 no lleva insert posicion 30 alfa sierra elemento alfa golf 35 no lleva insert posicion 30 alfa romeo elemento alfa golf 17 no lleva insert posicion 30 alfa quebec elemento alfa eco 03 no lleva insert posición, …"

In order to measure the WER, 2-min fragments of 6 random audios from each of the five nuclear plants were selected and manually transcribed, for a total of 1 h of transcribed audio. In this sample, containing a total of 2369 reference words, Whisper-timestamped obtained a WER of 7.7 %. We analyzed the main sources of errors in this sample. Some illustrative examples of this analysis are shown in Table 2. As can be seen in the table, the main problems we have encountered are as follows.

- Words in a language other than the base language, Spanish has been selected as the base language, but some of the words used in the task correspond to the NATO phonetic alphabet, and they are in English.
- Context problems, the model used (Whisper) has a feature that is to predict the word of the next time instant depending on the context, which sometimes produces transcription errors, transcribing a word that clearly does not correspond to what is heard in the audio.

As can be seen in Table 2, most of the transcription errors are related to the NATO phonetic alphabet and can be easily corrected by considering phonetically equivalent words in Spanish (alfa–alpha, eco-–echo, etc.) to obtain the codes. From the point of view of the application of this development, in which transcriptions are used to obtain the codes, the most important figure of merit is the accuracy in determining the codes, and this accuracy is very high, as shown in Table 3.

At the beginning of this development, several challenges were identified that were carefully always considered, the first being the size of the model used, playing a fundamental role in processing speed and accuracy. The second is how the models were going to behave in front of different speakers (accents, pronunciation …) and finally, the precision obtained for audios that differ from the purpose of the models used, since they are developed to translate or transcribe conversations, predict the next word before saying it, where the context has a lot of influence. In the case in question, audios have no context, a single speaker, with many periods of silent time and using a coding language and mixing Spanish and English.

The results obtained are considered exceptional, since in precision values are obtained that are considered acceptable for several reasons. Firstly, because it is a project developed with the feedback of the end client, and its specifications are met. Secondly, they are results analyzing samples from each of the facilities in which it operates, that is, results have been obtained from all possible cases. Thirdly, the samples from each facility analyzed consist of approximately 108,000s of audio, which makes a total of 540,000s analyzed.

This application has been developed under the VS-Code programming environment, in Python language and the Pytorch, Whisper-timestamp and Pyaudio libraries. This system run in computer with GPU NVIDIA RTX 4080. If used for post-processing, the model used is "large" whose average processing time for ~900s audios is ~150s. If used for real time, the model used is "small" and each ~30s fragment is processed in an average time of ~2s.

Table 2
Analysis of five examples of 2-min audio fragments, comparing reference with transcription.

EXAMPLE	WER	INSERTIONS	DELETED	SUBSTITUTED
1	0,053	–	–	5x(echo - eco)
2	0,000	–	–	–
3	0,259	posicion	–	11x(alpha - alfa) ', 1x(echo - eco), 1x(Charlie - charly)
4	0,026	–	delta	1x(kiko - kilo)
5	0,098	0		7x(echo - eco), 3x(whiskey -whisky)

Table 3
Percentage of correctly recognized codes from the audio transcription of the videos of different nuclear facilities. Each nuclear facility includes about 1000 fuel elements. Last column indicates the mean and standard deviation of the differences between the starting time of the code detected from the audio and from the video.

FACILITIES SET	ACCURACY (%)	μ /σ (s)
NPP-1	99.6	6.7/17.6
NPP-2	98.7	−1.4/39.5
NPP-3	99.5	2.6/12.8
NPP-4	99.3	0.0/5.8
NPP-5	98.6	0.2/41.1

5. Future work

Real-time streaming mode is useful in the situation at hand, since the analysis of the audio signal must be done simultaneously with the video signal, for live captioning. It means that the source speech audio has to be processed at the same time as it is recorded. The transcripts have to be delivered within a short additive latency, for example 1–3 s and the video signal must show the code at the same time.

Regarding audio, there are some implementations of Whisper for streaming, but their approach is rather naive, like first record a 30-s audio segment, and then processing it. The latency of these methods is large, and the quality on the segment boundaries is low because a simple content unaware segmentation can split a word in the middle. Another option is using the simple but effective Local Agreement algorithm [22], which is one streaming policy that can be used to convert any full-sequence to full-sequence model to operate in simultaneous streaming mode. Such an implementation is Whisper-Streaming [23].

During the review process of this article, work has been done on the real-time processing part. The starting point was very high precision but using the largest model (large), and therefore having a processing time of approximately 5s per 30s audio fragment.

Work has been done to reduce this time and finally a smaller model, (the 'small' model), has been chosen, along with some modifications to the parameters of the model itself, as well as the development of a real-time processing algorithm. With this, it has been possible to reduce transcription time to approximately 1–2s per 30s audio fragment and without reducing precision.

6. Conclusion

In this article, we propose and explore a novel step in the digitization of the process of obtaining the pool map of a nuclear power plant, in which we use the audio signal from the operator's microphone to obtain the identification codes of those components that are in each of the cells of the SFP. Audio processing provides us with an automatic acquisition system of the codes and a verification system to use in combination with the processing of the video signal.

We have developed an algorithm which uses at its core one of the latest models of multilingual ASR to transcribe audio signal. We build the identification code of fuel heads and other components with a post-processing of the transcriptions of the segments. Results show a very high level of accuracy in audio from stored videos. Besides, the methodology proposed is easily extensible to other nuclear facilities in the world because the model allows several languages.

This work presents a different and promising way of performing tasks in nuclear power plants by applying novel techniques such as Artificial Intelligence. The digitization of processes not only make it possible to avoid human failures when performing the task, but also reduce the number of people needed to carry out the task, thus reducing the exposure to possible radiation doses they may receive.

This effort is included in a large-scale project, in processing the audio signal, but there is still room for improvement, such as the possibility of using the model to identify codes of other fuel element components, as well as extending it to other similar processes.

Finally, it is worth mentioning the great novelty and importance of using advanced technologies in the nuclear field, something that is gradually being achieved. This project is a clear example of this trend and contributes to take a step forward, providing robustness and improving safety in the processes carried out in nuclear facilities.

CRediT authorship contribution statement

Sergio Segovia: Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Ángel Ramos:** Data curation, Formal analysis, Supervision, Validation. **David Izard:** Validation. **Doroteo T. Toledano:** Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: SERGIO SEGOVIA GONZALEZ reports financial support was provided by ENUSA Industrias Avanzadas, S.A, S.M.E. ANGEL RAMOS GALLARDO reports financial support was provided by ENUSA Industrias Avanzadas, S.A, S.M.E. DAVID IZARD HERNANDEZ reports financial support was provided by ENUSA Industrias Avanzadas, S.A, S.M.E. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by a ENUSA Industrias Avanzadas during the execution of the research project on the digitization of the process of obtaining the SFP map in nuclear power plant services. This work has also been partially funded by project PID2021-125943OB-I00 (Spanish Ministry of Science and Innovation and ERDF).

References

[1] IAEA, Classification of Radioactive Waste, 2009.

[2] IAEA, Data Requirements and Maintenance of Records for Spent Fuel Management: A Review, 2006.

[3] CSN, GS 01-07 Revisión 2 - Información a Remitir al CSN Por Los Titulares Sobre La Explotación de Las Centrales Nucleares, 2003.

[4] N. Shoman, P. Honnold, H. Smartt, D. Hannasch, Inspecta 1.0: implementation challenges for on-device speech and vision tasks, in: Proc. F the INMM/ESARDA Joint Annual Meeting, May 22-26, 2023, 2023.

[5] C. Tang, C. Yu, Y. Gao, J. Chen, J. Yang, J. Lang, J. Lv, Deep learning in nuclear industry: a survey, Big Data Mining and Analytics 5 (2) (2022) 140–160.

[6] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, IEEE Signal Process. Mag. 29 (2012) 82–97, https://doi.org/10.1109/msp.2012.2205597.

[7] R. Collobert, C. Puhrsch, G. Synnaeve, Wav2Letter: an End-To-End Convnet-Based Speech Recognition System, 2016 arXiv.Org, https://arxiv.org/abs/1609.03193.

[8] R. Prabhavalkar, K. Rao, T.N. Sainath, B. Li, L. Johnson, N. Jaitly, A comparison of sequence-to-sequence models for speech recognition, in: Interspeech 2017, ISCA, ISCA, 2017, https://doi.org/10.21437/interspeech.2017-233. (Accessed 31 October 2023).

[9] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig, Achieving Human Parity in Conversational Speech Recognition, 2016 arXiv.Org, https://arxiv.org/abs/1610.05256.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[11] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional Transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020.

[13] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: a framework for self-supervised learning of speech representations, Adv. Neural Inf. Process. Syst. 33 (2020) 12449–12460.

[14] J. Grosman, Fine-Tuned XLSR-53 Large Model for Speech Recognition in English, 2021. Available at: https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english. (Accessed 16 April 2024).

[15] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli, Facebook's Wav2Vec2, 2020. Available at: https://huggingface.co/facebook/wav2vec2-base-960h. (Accessed 16 April 2024).

[16] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, July, pp. 28492–28518.

[17] Gemini Team, Google, Gemini: A Family of Highly Capable Multimodal Models, 2024. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf. (Accessed 17 April 2024).

[18] OpenAI, Whisper-Large-v3, 2024. Available at: https://huggingface.co/openai/whisper-large-v3. (Accessed 17 April 2024).

[19] M. Bain, J. Huh, T. Han, A. Zisserman, WhisperX: Time-Accurate Speech Transcription of Long-form Audio, 2023 arXiv.Org, https://arxiv.org/abs/2303.00747.

[20] T. Giorgino, Computing and visualizing dynamic time warping alignments in R: the dtw package, J. Stat. Software 31 (2009), https://doi.org/10.18637/jss.v031.i07.

[21] J. Louradour, Whisper-timestamped: multilingual automatic speech recognition with word-level timestamps and confidence, GitHub Repository (2023). https://github.com/linto-ai/whisper-timestamped. Feb. 2024.

[22] D. Liu, G. Spanakis, J. Niehues, Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection, in: Interspeech 2020, 2020, pp. 3620–3624.

[23] D. Macháček, R. Dabre, O. Bojar, Turning whisper into real-time transcription system, in: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations, 2023, November, pp. 17–24.