Article

Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper

Juan Camilo Vasquez-Correa[1], **Aitor Alvarez-Muniain**[1]

[1] Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)

* Correspondence: {jcvasquez,aalvarez}@vicomtech.org (J.C.V., A.A.)

Abstract: The growth in online child exploitation material is a significant challenge for European Law Enforcement Agencies (LEAs). One of the most important sources of such online information corresponds to audio material that needs to be analyzed to find evidence in a timely and practical manner. That is why LEAs require a next-generation AI-powered platform to process audio data from online sources. We propose the use of speech recognition and keyword spotting to transcribe audiovisual data and to detect the presence of keywords related to child abuse. The considered models are based on two of the most accurate neural-based architectures to date: Wav2vec2.0 and Whisper. The systems are tested under an extensive set of scenarios in different languages. Additionally, keeping in mind that obtaining data from LEAs is very sensitive, we explore the use of federated learning to have more robust systems for the addressed application, while maintaining the privacy of the data to LEAs. The considered models achieved a word error rate between 11% and 25%, depending on the language. In addition, the systems are able to recognize a set of spotted words with true positives rates between 82% and 98%, depending on the language. Finally, federated learning strategies show that they can maintain and even improve the performance of the systems when compared to centralized trained models. The proposed systems sit the basis for an AI-powered platform for automatic analysis of audio in the context of forensic applications within child abuse. The use of federated learning is also promising for the addressed scenario, where data privacy is an important issue to be managed.

Keywords: Speech Recognition; Keyword Spotting; Child abuse; Federated Learning; Whisper; Wav2vec2.0

1. Introduction

The growth in online child exploitation and abuse material is a significant challenge for European Law Enforcement Agencies (LEAs). Currently, the revision of online material about child abuse exceeds the capacity of LEAs to respond in a practical and timely manner. One of the most important sources of information that needs to be analyzed to find evidence about child abuse corresponds to audiovisual material from multimedia content. With the aim to safeguard victims, prosecute offenders and limit the spread of online child abuse related material, LEAs need a next-generation AI-powered platform to process multimedia data from online sources. One of the main goals of the GRACE project[1] is to develop robust AI-based technology to equip LEAs with the aforementioned platform. Two of the core applications to be incorporated correspond to automatic speech recognition (ASR) and keyword spotting (KWS) in order to accurately transcribe audiovisual online material, and to detect the presence of specific keywords about child abuse in the transcriptions.

Within this context, ASR technology has been applied in different forensic scenarios. For instance, to collect evidence via the examination of electronic devices [1], or to analyze

[1] https://www.grace-fct.eu/

 doi:10.20944/preprints202212.0426.v1

multimedia content related to specific threats [2,3]. Nevertheless, the successful implementation of an ASR system in forensics introduces a series of issues to be solved, which are not present in other domains where ASR is applied. For instance, it is common to find audio coming from different sources, which are highly affected by background noise, overlapping speakers, audio reverberation, among other factors. All these aspects affect the quality of the obtained transcription and the capability of the system to detect specific keywords.

Although the aforementioned problems, recent advances in ASR have introduced novel end-to-end architectures [4] that have shown to be accurate enough in those adverse conditions. The core idea of end-to-end models is to directly map the input speech signal to character sequences and therefore greatly simplify training, fine-tuning and inference [5–9]. Two main approaches are distinguished in the literature to train end-to-end ASR systems: fully supervised or self-supervised models. Regarding the first group, NVIDIA proposed Quartznet [10] with the aim to build a competitive but lighter end-to-end ASR model. The architecture consists of multiple blocks of 1D convolutions stacked with residual connections. The model has been trained and tested on the Common Voice corpus, achieving Word Error Rates (WERs) between 7.7% and 12.5%, depending on the language [11]. A Quartznet model also produced WERs of 19.2% and 18.3% in French and Spanish language multimedia data, respectively, from the MediaSpeech corpus [12]. Researchers from NVIDIA recently proposed Citrinet [13] as an evolution of Quartznet. The model consists of a residual network formed by 1D time-channel separable convolutions combined with a sub-word encoding and a squeeze-and-excitation mechanism [14]. The authors reported a WER of 5.6% on the TEDLIUMv2 corpus. Another architecture that have proven to be accurate in many ASR benchmark scenarios is the Recurrent Neural Network Transducer (RNN-T) [15]. The RNN-T is formed by three main blocks: (1) an encoder network that receives input acoustic frames and produces high-level speech representations, (2) a predictor that acts as a decoder by processing the previous produced token, and (3) a joint network that combines the outputs from the two previous blocks and produces the distribution of the next predicted token or *blank* symbol. Recent models based on RNN-T achieved a WER 14.0% in the TEDLIUMv2 corpus [16].

Contrary to fully supervised models, recent studies are focused on the use of big acoustic models trained with self-supervised learning methods and a large amount of unlabelled data. Researchers from Meta AI demonstrated the capabilities of this type of models by introducing Wav2Vec2.0 [17]. This system outperformed many benchmark results, especially when considering ASR for low-resource languages in the Common Voice corpus [18]. Particularly, the authors in [19] considered a Wav2vec2.0 model combined with their proposed language modeling approach, and achieve state-of-the-art results in the German Common Voice corpus, with a WER of 3.7%. Wav2Vec2.0-based models have also been successfully tested in more adverse acoustic environments such as in multimedia Portuguese data from the CORAA database [20]. Due to these reasons, Wav2Vec2.0 has become one of the most considered neural-based models for ASR. Self-supervised approaches like Wav2Vec2.0 are challenging because there is not a predefined lexicon for the input sound units during the pre-training phase. Moreover, sound units have variable length with no explicit segmentation [21]. With the aim to solve such issues, Meta AI released HuBERT as a new approach to learn self-supervised speech representations [22]. The combination of Convolutional and Transfomer networks from Wav2Vec2.0 and HuBERT has achieved state-of-the-art results in many ASR scenarios. With the aim to combine the best features from both type of networks in a single neural block, researchers from Google introduced the "convolutional augmented Transfomer" or Conformer [23]. A Conformer network achieved a WER of 7.2% in the TEDLIUMv2 corpus [24].

Self-supervised audio encoders like Wav2Vec2.0, HuBERT, and Conformers learn high quality audio representations. However, due to its unsupervised pre-training nature, they lack a proper decoding to transform such representations into usable outputs. This is why a fine-tuning stage is always necessary in order to accurately implement models for ASR or audio classification. With the aim to solve the aforementioned issue, researchers from

 doi:10.20944/preprints202212.0426.v1

OpenAI recently proposed "Whisper" [25]. Whisper is a sequence-to-sequence Transformer trained in a fully supervised manner, using up to 680,000 hours of labelled audio from internet. The model has achieved state-of-the-art WER results in many benchmark datasets for ASR, including librispeech, TEDLIUM, Common Voice, among others.

There are two main issues that appear when designing ASR solutions for forensic scenarios: The first one is related to find the most appropriate neural architecture from the ones previously described in order to deal with different acoustic environments. The second one is related to data privacy and protection [26]. Generally, obtaining operative data from LEAs for the addressed scenario is not possible. In this context, Federated Learning (FL) has emerged as an alternative to train machine learning models over remote devices such as mobile phones or remote data-centers in a non-centralized manner, preserving data privacy [27–30]. The procedure is as follows: LEAs operative data are stored in on-premise data servers. Then, FL strategies aim to transfer only local model updates to a central server, keeping LEAs data private. The central server aggregates information obtained from multiple clients i.e., LEAs, and updates a central model that is transmitted back to the clients for their consumption. FL has been applied to train robust federated acoustic models for ASR [31–33] and KWS [34]. In [32] the authors proposed a client adaptive federated training to mitigate data heterogeneity when training ASR models. The proposed system achieved a similar WER with respect to the obtained one using a fully centralized training. In [33] the authors proposed a strategy to compensate Non Independent and Identically Distributed (non-IID) data in federated training of ASR systems. The proposed strategy involved random client data sampling, which resulted in a cost-quality trade-off. The optimization of such a trade-off led to obtain ASRs with similar WERs than the obtained by training centralized systems. The authors in [34] demonstrated the capabilities of federated training to obtain robust KWS systems locally trained on edge devices like smartphones, reaching similar accuracies when compared with centralized trained models.

According to the reviewed literature, the two main paradigms and solutions for ASR to date include self-supervised models based on Wav2Vec2.0 and fully supervised models such as Whisper. This work considered and compared these two approaches to test their capabilities to perform robust ASR and KWS in a large set of test scenarios. We also evaluated the use of FL in the context where different LEAs can share a common ASR and KWS system keeping the privacy of their data. In summary, the main contributions of this paper are four-folds:

1. We performed an extensive comparison between two of the most accurate neural-based ASR architectures to date: a fine-tuned version of Wav2Vec2.0 and Whisper. The evaluation is performed in many scenarios, but paying special attention to corpora coming from multimedia content. The models are tested in data from seven indo-European Languages, including English, Spanish, German, French, Italian, Portuguese, and Polish. This evaluation can be useful as well in other domains besides ASR forensics, making our contribution open and viable in other scenarios.
2. We created and released an in domain corpus that includes specific keywords of child abuse domain, and a set of accompanying audios where the keywords are present. The included audios are selected from open available corpora used in the literature. The created corpus can be used as a benchmark to test ASRs in non-controlled acoustic conditions.
3. The two neural architectures are compared as well in the created corpora within the scope of child abuse forensics. To the best of our knowledge, this is the first study that comprises the use of open ASR solutions and their capabilities to recognize specific words within a forensic domain.
4. We validated the use of FL strategies to train ASR systems in the context of forensic applications. The core idea is that different LEAs can share a common model but keeping the privacy of their data.

The rest of the paper is distributed as follows. Section 2 details different technical aspects of Wav2Vec2.0 and Whisper architectures for ASR. Section 3 describes the consid-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0426.v1

ered corpora to test the ASR systems, and the process to deliver an in domain corpus for KWS in the context of forensics. Section 4 describes the pilot study on the use of FL for the addressed application. Section 5 displays the main results obtained regarding ASR, KWS, and FL. Section 6 discusses the main insights obtained from the results. Finally, Section 7 shows the main conclusion derived from this work.

2. Methods

We considered two of the most accurate neural-based ASR architectures to date: (1) Wav2vec2.0, which is trained following a self-supervised paradigm, and (2) Whisper, which is trained following a fully supervised strategy. Details about each model are found in the following sub-sections.

2.1. Wav2vec2.0

Wav2vec2.0 [17] is a self-supervised end-to-end architecture based on convolutional and Transformer layers (see Figure 1). The model encodes raw audio waveforms χ into latent speech representations $z_1, \ldots, z_T$ via a multi-layer convolutional feature encoder $f : \chi \rightarrow Z$. These latent representations fed a Transformer-masked network $g : Z \rightarrow C$. The Transformer network initially quantise the continuous representations, forming a discrete set of outputs $q_1, \ldots, q_T$ that represent targets in the self-supervised learning objective [17, 35]. Those quantised representations are then contextualised using the attention blocks from the Transformer module, obtaining a set of discrete contextual representations $c_1, \ldots, c_T$. The feature encoder is formed by seven convolutional blocks with 512 channels, strides of $\{5, 2, 2, 2, 2, 2, 2\}$ and kernel widths of $\{10, 3, 3, 3, 3, 2, 2\}$. The Transformer network is formed by 24 blocks, a model dimension of 1024, an inner dimension of 4096 and a total of 16 attention heads.

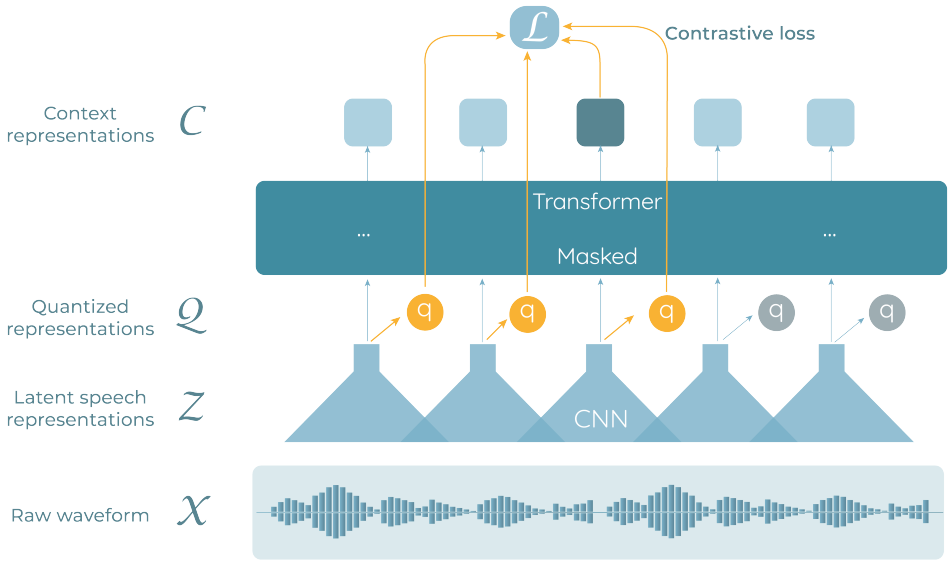

Figure 1. Wav2vec2.0 architecture representation. The raw audio signal is mapped to speech representations that are fed into a Transformer network to output context representations. Figure based on the one presented in [17].

We considered a pre-trained Wav2vec2.0 acoustic model based on the Wav2Vec2-XLS-R-300M model, which is available via Hugginface[2]. The model was pre-trained in a self-supervised manner using 436k hours of unlabelled speech data in 128 languages

[2] https://huggingface.co/facebook/wav2vec2-xls-r-300m

 doi:10.20944/preprints202212.0426.v1

from the VoxPopuli [36], Multilingual librispeech (MLS) [37], Common Voice [38], BABEL, and VoxLingua107 [39] corpora. The Wav2Vec2-XLS-R-300M is one of the different versions of the Meta AI's XLS-R multilingual model [40] composed by 300 million parameters. The multilingual pre-trained model was fine-tuned with labelled speech data (see Section 3.1) in seven languages: English, German, French, Spanish, Italian, Portuguese, and Polish. Each model was trained for 50 epochs, with a batch size of 2, 16 gradient accumulation steps, and a learning rate of 5×10^{-5}, which is warmed up during the initial 10% of the training.

The trained acoustic representations are decoded using a Connectionist Temporal Classification (CTC) layer with a beam-search decoding strategy (beam-width=256). The CTC decoding include the use of separate 3-gram language models that are trained using large text corpora, and which are included in the decoding with weights of $\alpha = 0.5$ and $\beta = 1.5$.

2.2. Whisper

Whisper is a recently introduced ASR system by OpenAI [25]. Contrary to Wav2vec2.0, Whisper is trained in a fully supervised manner, using up to 680k hours of labelled speech data from multiple sources. The model is based on an encoder-decoder Transformer, which is fed by 80-channel log-Mel spectrograms. The encoder is formed by two convolution layers with a kernel size of 3, followed by a sinusoidal positional encoding, and a stacked set of Transformer blocks. The decoder uses the learned positional embeddings and the same number of Transformer blocks from the encoder. Figure 2 illustrates the general Whisper architecture. Different pre-trained models are available with variations in the number of layers and attention heads. We considered the "Whisper-large" model, which consists of 1550 million parameter distributed in 32 layers and 20 attention heads. The model is available via Huggingface[3].

[3] https://huggingface.co/openai/whisper-large

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0426.v1

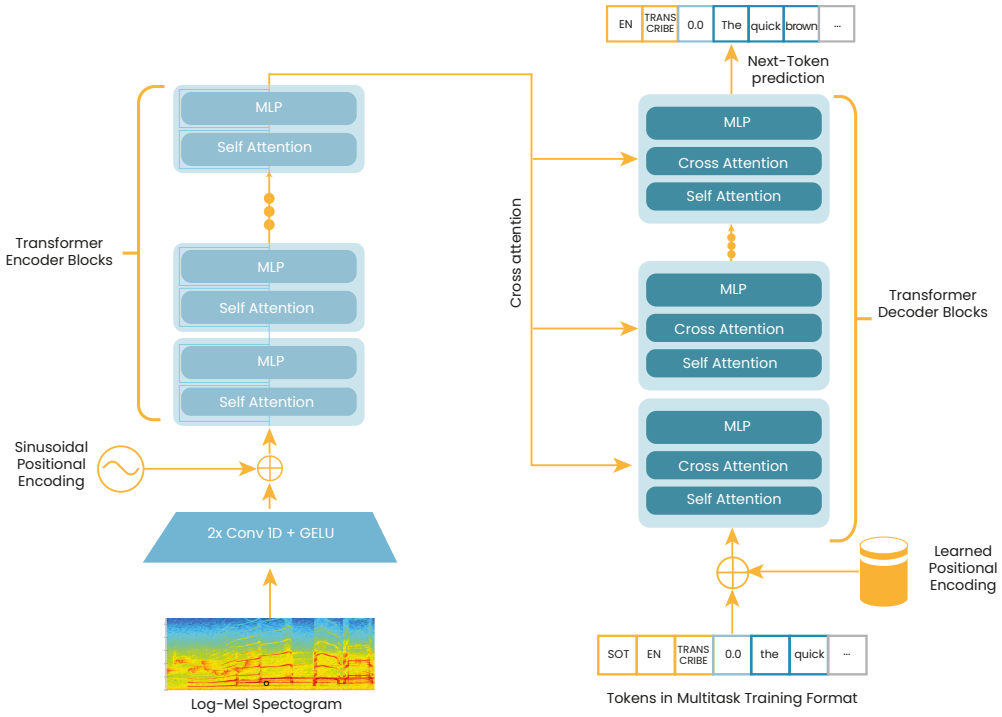

Figure 2. Whisper architecture representation. The log Mel-spectrograms are encoded by a Transformer network. Encoded representations are transformed into character outputs and no-speech tokens via the Transformer decoder. Figure based on the one presented in [25].

The model was not fine-tuned in this study, thus the evaluation for all languages was conducted in a zero-shot setting. The decoding was performed using a beam search strategy with 5 beams, an array of temperature weights of $[0.2, 0.4, 0.6, 0.8, 1]$, and a no repeat n-gram size of 3 in order to take advantage of the language modeling head and to avoid loops, in a similar way to [25].

3. Materials

This section describes a set of open corpora used to benchmark the two considered ASR systems (Section 3.1), followed by the performed process to derive a set of keywords to be spotted by the considered systems (Section 3.2), and the description of a built in domain dataset considered as well to test the considered models (Section 3.3).

 doi:10.20944/preprints202212.0426.v1

3.1. Data

Table 1. List of public speech corpora considered to test the performance of ASR and KWS systems based on Wav2Vec2.0 and Whisper.

Corpus name	Description	Languages	Test Duration (h)
Common Voice [38]	Read sentences collected and validated via crowd-sourcing	English German French Spanish Italian Portuguese Polish	173 72 38 26 23 6 7
Spoken Wikipedia Corpus (SWC) [41]	Volunteer readers of Wikipedia articles	English German	42 36
Media Speech [12]	Speech segments from YouTube videos	French Spanish	10 10
Multilingual TEDx [42]	Audio recordings and transcripts from TED talks	German French Spanish Italian Portuguese	2 2 2 2 2
TEDLIUMv2 [43]	Audio recordings from TED talks	English	3
Multilingual librispeech (MLS) [37]	Audio recordings from audiobooks	German French Spanish Italian Polish Portuguese	14 10 10 5 2 4
Voxforge	Crowdsourced read speech	German French Spanish Italian Portuguese	3 4 5 2 1
Debating technologies [44]	Audio recordings from transcribed public debates	English	1
Polish Parliamentary corpus [45]	Recordings from the Polish parliament	Polish	1
CORAA [20]	Combination of five corpora in Portuguese	Portuguese	13

Different public corpora were considered to train/test the ASR and KWS models. Wav2vec2.0 models were fine-tuned using the Common Voice corpus [38] for each considered language. The amount of available labelled data highly varies depending on the language, and include: 1600 hours for English, 777 hours for German, 623 for French, 324 for Spanish, 158 for Italian, 63 for Portuguese, and 43 for Polish. These data are freely available via Huggingface[4]. The training data for the Spanish model also included 57 hours from the RTVE2018 dataset [46] from the Albayzin 2018 evaluation challenge.

The performance of both the fine-tuned Wav2vec2.0 and Whisper-based models was evaluated in a cross-corpora fashion, considering a large set of databases from the literature that are available in the different languages. The list of considered corpora is observed

[4] https://huggingface.co/datasets/common_voice

 doi:10.20944/preprints202212.0426.v1

in Table 1. These corpora were selected in order to test the performance of the models in several recording conditions, which can be closer to the realistic scenarios found by LEAs. Notice that due to the sensitive nature of the target application, it is not possible to get access to realistic operative data from LEAs. However, we created an in-domain synthetic dataset using these open source corpora, which is described in Section 3.3.

3.2. Spotted keywords

In order to test the capabilities of the ASR models to spot specific keywords within the child abuse domain, we defined a list of keywords to be spotted. The keyword list was obtained from a set of open documents that include: (1) the "Best Practices on Victim support for LEA first responders" deliverable from the GRACE project[5], (2) the 2021 "Barriers to Compensation for Child Victims of Sexual Exploitation" report from ECPAT[6] [47], (3) the study from [48], (4) EUROPOL technical reports [49–51], (5) EUROPOL press-releases from 2018 to 2022 using the keyword "child abuse"[7], (6) Wikipedia articles about "child abuse" and "online child abuse", and (7) UNICEF press-releases about "child abuse"[8]. All documents were text crawled and pre-processed by performing lemmatisation, and removing stop words, numbers, and date entities. After this process, we obtained a corpus with 55,059 words, whose 6028 are unique. Figure 3 shows the most important keywords found in the crawled corpus.

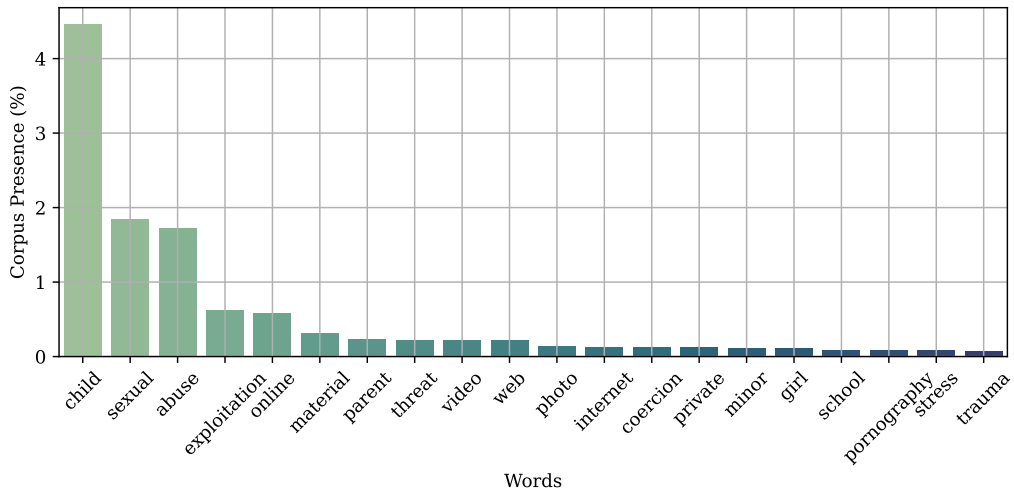

Figure 3. Top 20 of the most important keywords related to child abuse, which were used to test the capability of the ASR system to detect specific terminology within the domain.

Afterwards, we selected the 100 most repeated words from the corpus, which represent the 33% of the information within the whole set of crawled documents. Finally, we excluded 12 terms because they were very broad concepts not related with child abuse, leading to a final set of 88 keywords to be spotted. The obtained keyword list (in English) was translated into the remaining six considered languages in order to have a common benchmark for all languages.

3.3. GRACE dataset

We considered an additional corpus to test the implemented ASR systems by merging and filtering the data described in Section 3.1. We selected audio samples from all datasets

[5] https://www.grace-fct.eu/deliverables/70
[6] https://ecpat.org/wp-content/uploads/2021/05/Barriers-to-Compensation-for-Child_ebook.pdf
[7] https://www.europol.europa.eu/media-press/newsroom?q=child%20abuse
[8] https://www.unicef.org/search?force=0&query=child+abuse&created%5Bmin%5D=&created%5Bmax%5D=

 doi:10.20944/preprints202212.0426.v1

that contain at least one of the 88 selected keywords. Table 2 shows the data distribution for each language after selection. The table includes the datasets considered for each language where the keywords are found, the number of utterances, and the total audio duration (in hours).

Table 2. Data distribution for the GRACE dataset, which combines different corpora into a single one within the child abuse domain.

Language	Base corpora	# utterances	Duration (h)
English	SWC, Debating technologies, TEDLIUMv2	2979	9.2
German	Multilingual TEDx, SWC, Voxforge	1712	5.9
French	Multilingual TEDx, MediaSpeech, Voxforge	1250	4.1
Spanish	Multilingual TEDx, MediaSpeech, Voxforge	557	2.0
Italian	Multilingual TEDx, Voxforge	354	1.0
Portuguese	Multilingual TEDx, Voxforge, CORAA	1503	2.3

The selected audios were processed in order to have also more realistic acoustic conditions than those expected in forensic applications within the considered domain. The process includes: (1) adding background noise with signal to noise ratios (SNR) between 5 and 30 dB (randomly), (2) adding reverberation using room impulses from the VOiCES dataset [52], and (3) randomly applying the ogg-vorbis codec [53] due to it is commonly found in audio material from online sources. The final ASR and KWS evaluation is performed considering the two versions of the corpus: clean and noisy. This corpus is available online[9] to be used as a benchmark dataset for speech recognition in different languages under non-controlled acoustic conditions.

4. Federated Learning

The considered FL pipeline is performed only with English data and includes five nodes that are used for federated training, a dummy node considered to test the evolution of the learning process, and the central server in charge of aggregating the weights received from the five nodes. Figure 4 shows the implemented architecture. Three of the servers were located at Vicomtech premises (Spain), one server was located at Greece, another one in Portugal, and the remaining one in Cyprus. The aim of these connections is to create a real environment for the pilot, in similar conditions to the expected when the model is trained by different LEAs across Europe. In addition, secure communication between clients and the server was established through a VPN connection to ensure that sensitive data (parameters) are safely transmitted and to prevent unauthorised access. Each node contains data from a different dataset: TEDLIUMv2, debating technologies, Librispeech-other, Librispeech-clean, and SWC. This data configuration aims to evaluate the impact of non-IID data distribution, which is more realistic for the addressed forensic application.

[9] https://datasets.vicomtech.org/di01-grace-automatic-speech-recognition-and-keyword-spotting/GRACE_ASR.zip

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0426.v1

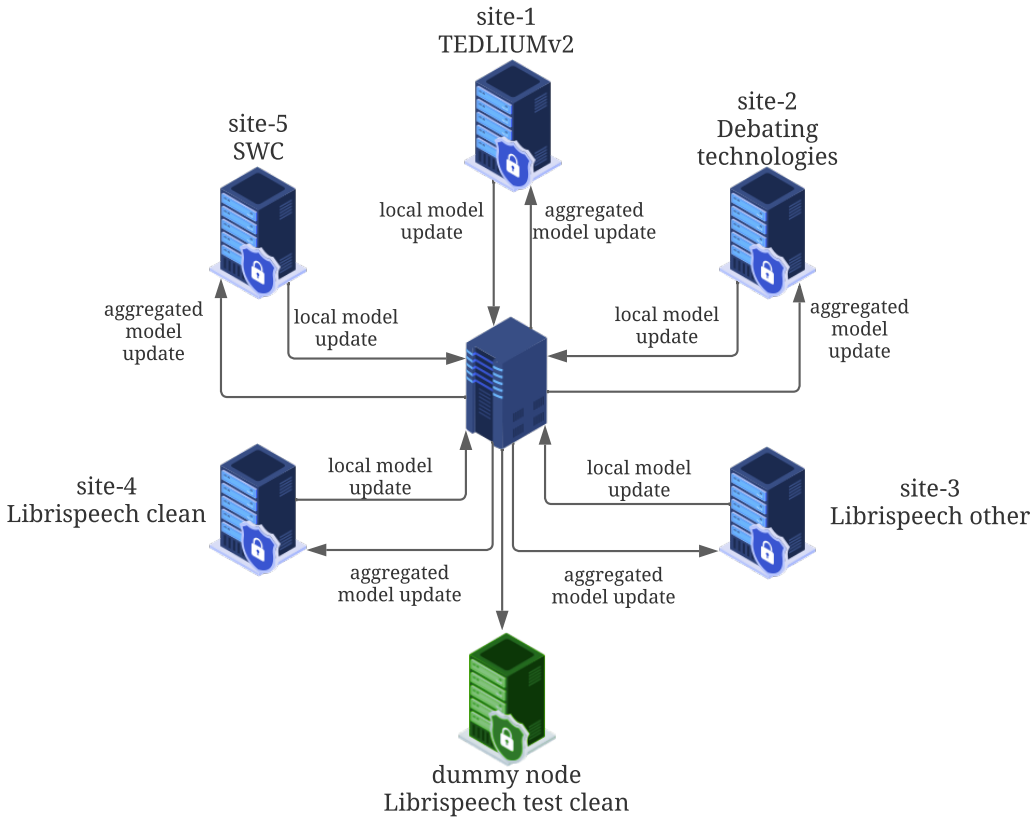

Figure 4. Configuration of the FL architecture. Central server with five client nodes (site-$\{1, 2, \cdots, 5\}$) and a dummy node only used to test the performance of the aggregated model

The FL pilot was performed only with the Wav2Vec2.0 model, and using also the pre-trained Wav2Vec2-XLS-R-300M model. The training hyperparameters were the same for the five clients, and include a batch size of 2, a learning rate of 5×10^{-5} warmed up in the first 10% of the training time, and a gradient accumulation of 16 steps. The local training is performed for 5 epochs. The central server is configured to run for 10 rounds of federated training, and using the Federated averaging (FedAvg) aggregation mechanism to update the central model. The architecture configuration and the training process is implemented using Nvidia Flare[10].

5. Results

5.1. Speech Recognition

Wav2Vec2.0 and Whisper models were evaluated under the described corpora in Section 3.1. The results of the ASR systems in terms of WER are shown in Table 3. The results included those obtained in the evaluation of the seven languages, and using both the open benchmark corpora and the two versions (clean and noisy) of the synthetic GRACE corpus.

[10] https://nvflare.readthedocs.io/en/main/index.html

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0426.v1

Table 3. Results of the ASR models in different languages considering all benchmark datasets. Results in terms of WER.

Model	Common Voice	MLS	TED-LIUMv2	MTEDx	SWC	Media Speech	Voxforge	Debates	Polish Parl	CORAA	GRACE clean	GRACE noisy	AVG.
English													
Wav2Vec 2.0	16.1	-	17.2	-	20.6	-	-	11.7	-	-	**18.9**	32.6	19.5
Whisper	**10.0**	-	**5.4**	-	20.6	-	-	**7.0**	-	-	24.5	**19.8**	**14.6**
German													
Wav2Vec 2.0	11.9	12.9	-	36.7	34.5	-	7.5	-	-	-	20.0	33.8	22.5
Whisper	**7.1**	**6.7**	-	**21.7**	**18.3**		**4.2**	-	-	-	**15.5**	**22.5**	**13.7**
French													
Wav2Vec 2.0	**16.7**	17.0	-	25.3	-	**29.1**	16.7	-	-	-	**26.5**	56.3	26.8
Whisper	21.7	**8.0**	-	**23.3**	-	35.8	**14.6**	-	-	-	36.8	**34.1**	**24.9**
Spanish													
Wav2Vec 2.0	**4.7**	7.2	-	12.9	-	**14.5**	6.3	-	-	-	**12.6**	33.3	13.1
Whisper	6.2	**5.3**	-	**9.4**	-	15.8	**4.2**	-	-	-	19.6	**18.8**	**11.3**
Italian													
Wav2Vec 2.0	12.8	21.1	-	22.2	-	-	14.3	-	-	-	18.0	46.3	22.5
Whisper	**7.9**	**13.6**	-	**11.6**	-	-	**10.5**	-	-	-	**14.1**	**20.2**	**13.0**
Portuguese													
Wav2Vec 2.0	12.9	20.1	-	33.8	-	-	17.8	-	-	48.5	42.7	68.1	34.8
Whisper	**5.4**	**8.8**	-	**13.1**	-	-	**11.2**	-	-	**21.7**	**22.1**	**42.3**	**17.8**
Polish													
Wav2Vec 2.0	11.5	12.7	-	-	-	-	-	-	**32.1**	-	-	-	18.8
Whisper	**8.9**	**6.0**	-	-	-	-	-	-	32.5	-	-	-	**15.8**

On average, the WER for each language using Whisper ranges from 11.3% (in Spanish) to 24.9% (in French). The results using Wav2Vec2.0 range from 13.1% (in Spanish) to 34.8% (in Portuguese). In general, Whisper produces less errors than Wav2Vec2.0 (see Figure 5 left). The difference between both models is statistically significant according to a Mann Whitney test (U=1203.5, p-value=0.016). Whisper outperformed Wav2Vec2.0 especially under the most affected acoustic conditions, such as in the GRACE noisy, TEDLIUMv2, Debates, and CORAA corpora. However, there are some scenarios where Wav2Vec2.0 outperformed Whisper and which should be considered with special attention, such as the results for Spanish Common Voice.

The results obtained were compared to those found in the literature for the multilingual corpora: Common Voice, MLS, MTEDx, and MediaSpeech. The comparison is shown in Table 4. The Wav2Vec2.0-based model outperformed results in the Spanish versions of Common Voice and MediaSpeech corpora, with WERs of 4.3% and 14.5%, respectively with respect to to the results reported in [18] for Common Voice (WER=6.2%) and in [12] for MediaSpeech (WER=18.3%). We also reported state-of-the-art results for the Spanish, Portuguese, Italian, and German versions of the MTEDx corpus (WERs of 9.4%, 12%, 11.6%, and 21.7%, respectively) with respect to the WERs of 16.2%, 20.2%, 16.4%, and 42.3% reported in [42]. Whisper model also achieved state-of-the-art results in the CORAA corpus (WER=21.7%) with respect to the results reported in [54] (WER=21.9%), and in the TEDLIUMv2 corpus (WER=5.4%) compared to [13] (WER=5.6%). Regarding MLS, the state-of-the-art results are still from [55]. However, notice that the results reported here correspond to cross-corpus tests, while the experiments performed in [55] correspond to Wav2Vec2.0 models trained and tested using MLS, thus making the models adapted just for such a corpus.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0426.v1

Table 4. WER comparison between the results reported and those coming from the state-of-the-art for Common Voice, MLS, and MTEDx corpora. Best results for each corpus and language are highlighted in bold

Corpus	Reference	Language						
		English	German	French	Spanish	Italian	Portuguese	Polish
Common Voice	[11]	-	7.7	12.5	10.9	-	-	-
	[18]	-	7.2	11.2	6.2	**6.5**	6.1	**7.6**
	[36]	-	7.8	**9.6**	10.0	-	-	-
	[25]	10.1	7.7	14.7	6.4	8.1	7.1	9.0
	[19]	-	**3.6**	-	-	-	-	-
	[56]	-	9.8	-	-	-	-	-
	[57]	-	-	-	-	-	9.2	-
	Wav2vec2.0	16.1	11.9	16.7	**4.7**	12.8	12.9	11.5
	Whisper-large	**10.0**	7.1	21.7	6.2	7.9	**5.4**	8.9
MLS	[37]	-	6.5	5.6	6.1	10.5	19.5	20.4
	[40]	-	7.4	10.0	6.9	12.0	15.6	9.8
	[55]	-	**4.1**	**5.0**	**3.7**	**8.2**	**8.0**	**6.6**
	[25]	-	6.6	8.9	5.4	14.3	9.2	**6.6**
	[57]	-	-	-	-	-	12.3	-
	Wav2vec2.0	-	12.9	17.0	7.2	21.1	20.1	12.7
	Whisper-large	-	6.7	8.0	5.3	15.8	8.8	9.9
MTEDx	[42]	-	42.3	**19.4**	16.2	16.4	20.2	-
	[57]	-	-	-	-	-	21.0	-
	Wav2vec2.0	-	36.7	25.3	12.9	22.2	33.8	-
	Whisper-large	-	**21.7**	23.3	**9.4**	**11.6**	**13.1**	-
MediaSpeech	[12]	-	**19.2**	18.3	-	-	-	-
	Wav2vec2.0	-	29.1	**14.5**	-	-	-	-
	Whisper-large	-	35.8	15.8	-	-	-	-

5.2. Keyword Spotting

The text transcriptions from Wav2Vec2.0 and Whisper were post-processed in order to find the presence of the defined keywords to be spotted. The process involved transforming the transcription to lowercase and lemmatization. Lemmatization is performed to reduce the inflectional form of each word in order to detect all possible variations of the word within the transcription. The lemmatization process is performed using the set of large open dictionaries available in Spacy[11]. The results obtained for KWS in each corpus are shown in Table 5. The results are presented in terms of the true positive rate (TPR). This is a common metric used in this type of applications where it is more important to avoid false positive than false negative errors [58,59].

On average, the TPRs are higher using Whisper, and the results per language using Whisper range from 81.5% (for Polish) to 98.4% (for Italian). Results using Wav2Vec2.0 range from 82.9% (for Portuguese) to 94.9% (for Spanish). Similar to the ASR results, the difference between Whisper and Wav2Vec2.0 is larger when considering speech signals in non-controlled acoustic conditions, like the ones from the GRACE noisy corpus, where we particularly guarantee the presence of the spotted keywords in every utterance. High differences were also observed int the CORAA corpus, in Common Voice, and in the German SWC. The difference between the results obtained using Wav2Vec2.0 and Whisper is also statistically significant (see Figure 5 right) according to a Mann-Whitney test with U=589.0 and a p-value=0.003.

[11] https://spacy.io/usage/models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2022 doi:10.20944/preprints202212.0426.v1

Table 5. Results of KWS in different languages considering all benchmark datasets. Results in terms of TPR (%).

Model	Common Voice	MLS	TED-LIUMv2	MTEDx	SWC	Media Speech	Voxforge	Debates	Polish Parl	CORAA	GRACE clean	GRACE noisy	AVG.
English													
Wav2Vec 2.0	93.3	-	95.4	-	92.5	-	-	96.6	-	-	**94.6**	79.5	92.0
Whisper	**96.8**	-	**97.4**	-	**94.1**	-	-	**97.7**	-	-	91.8	**93.6**	**95.2**
German													
Wav2Vec 2.0	91.3	96.9	-	93.5	80.4	-	99.8	-	-	-	94.3	79.7	90.8
Whisper	**97.8**	**98.8**	-	**97.7**	**97.6**		99.8	-	-	-	**97.2**	**90.6**	**97.1**
French													
Wav2Vec 2.0	90.6	90.1	-	**94.5**	-	82.8	90.7	-	-	-	**90.5**	60.1	85.6
Whisper	**94.6**	**98.0**	-	93.9	-	**84.9**	**94.2**	-	-	-	88.0	**81.3**	**90.7**
Spanish													
Wav2Vec 2.0	96.7	98.1	-	98.1	-	96.3	**100.0**	-	-	-	**97.0**	78.1	94.9
Whisper	**98.2**	**99.8**	-	**98.6**	-	**94.4**	99.8	-	-	-	92.0	**94.2**	**96.7**
Italian													
Wav2Vec 2.0	90.7	97.2	-	95.5	-	-	98.9	-	-	-	96.1	80.9	93.2
Whisper	**97.8**	**99.9**	-	**97.3**	-	-	**99.8**	-	-	-	**98.7**	**96.9**	**98.4**
Portuguese													
Wav2Vec 2.0	93.1	93.4	-	94.1	-	-	99.1	-	-	74.3	76.8	49.5	82.9
Whisper	**96.6**	**97.5**	-	**99.4**	-	-	**100.0**	-	-	**88.1**	**88.3**	**81.3**	**93.0**
Polish													
Wav2Vec 2.0	93.9	96.9	-	-	-	-	-	-	**83.3**	-	-	-	**91.4**
Whisper	**95.4**	**98.7**	-	-	-	-	-	-	50.3	-	-	-	81.5

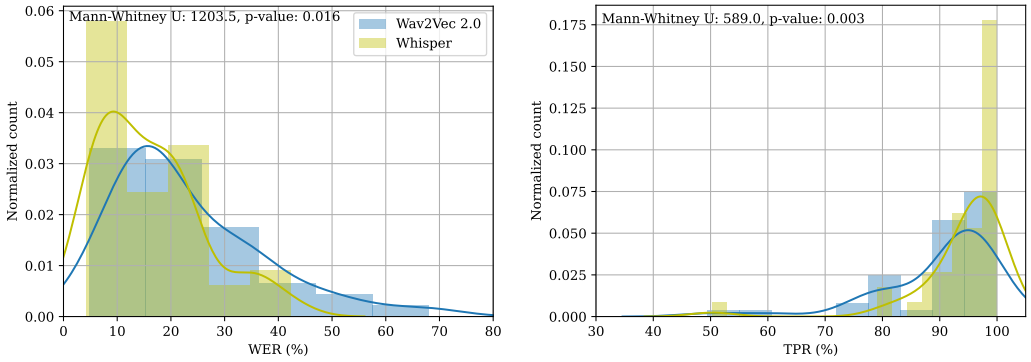

Figure 5. Comparison between the results obtained using Wav2Vec2.0 and Whisper for ASR (left) and KWS (right).

5.3. Federated Learning

The FL experiment involved training the Wav2Vec2.0 system using 5 separate real servers for training, and one additional node (dummy) used only to test the final model. Each node contained data from a different dataset (only in English) in order to evaluate the contribution from each corpus into the global aggregated model. The aim was also to cover non-IID conditions, which have shown to be one of the most important drawbacks when training models in an FL approach. The results are shown in Table 6. The results using the FL training are compared to those obtained training the system in a complete centralized manner. Similar WERs were obtained in each node comparing the federated vs. centralized training. The main difference is that when considering FL models there is only one aggregated model which covers the results of the 5 nodes, instead of having 5 different models for the case of the centralized approach. This fact highly reduces the time considered to train the system, and most important, it is possible to take advantage of data

 doi:10.20944/preprints202212.0426.v1

from different data centers to train a more robust and general model without the need of sharing data among clients.

Table 6. Results of the FL pilot comparing WERs from Wav2Vec2.0 models trained in a federated or centralized way.

Node	Data	WER Federated	WER centralized
node-1	TED-LIUMv2	13.5	13.3
node-2	Debates	12.4	12.3
node-3	Librispeech-other	7.9	7.8
node-4	Librispeech-clean	2.8	3.2
node-5	SWC	25.7	24.3
dummy	Librispeech-clean	3.6	3.8

6. Discussion

The evaluation of Wav2Vec2.0 and Whisper-based ASR systems is performed under a large set of different scenarios, including one specifically designed for forensic applications within child domain abuse. On average, Whisper is more accurate than the Wav2Vec2.0-based system. Whisper achieved WERs ranging from 11.5% to 24.9%, depending on the language, compared with Wav2Vec2.0 WERs that range between 13.3% to 34.8%. The difference between the two models is even larger when considering languages trained with lower resources, such as Portuguese or Italian. Although these differences, Wav2vec2.0 is competitive with Whisper when the number of hours for fine-tuning is large, e.g, in English, Spanish, or French.

Results using the GRACE dataset show relatively similar WERs between Wav2Vec2.0 and Whisper when considering the clean version of the corpus, with an average WER of 22.1% for Whisper and of 23.2% for Wav2Vec2.0. However, the difference between the two models greatly increases when considering the noisy version of the corpus, with an average WER of 26.3% for Whisper and of 45.1% for Wav2Vec2.0. This is a great indicator about the capability of Whisper to perform accurate transcriptions under non-controlled and noisy acoustic conditions, by keeping similar WERs in the two versions of the GRACE corpus. Despite the differences between the two types of models, there are some surprising results where Wav2Vec2.0 outperforms Whisper, and which should be considered with special attention. For instance, when evaluating the GRACE clean corpus in languages such as English, French, and Spanish. The models for these three languages were fine-tuned with more data, which likely explains the WER reduction in Wav2Vec2.0 with respect to Whisper.

The performed evaluations of our systems achieved state-of-the art results in several of the considered benchmark corpora. We reported state-of-the-art results for some of the languages in the Common Voice corpus. State-of-the-art results were also achieved for almost all languages in the MTEDx and MediaSpeech corpora. These results are good indicators about the capabilities of the considered systems to accurately recognize speech under more natural and spontaneous scenarios, closer to the expected in forensic domains.

The KWS evaluation indicated that both Wav2Vec2.0 and Whisper were accurate enough to recognize the considered child abuse-related keywords in the seven languages. TPRs obtained for Wav2Vec2.0 range from 82.9% to 94.9%, depending on the language. Results using Whisper range from 80.3% to 98.2%. The particular evaluation of KWS in the GRACE dataset also shows that both models are equally accurate to recognize the selected keywords under controlled acoustic conditions. On the contrary, when considering the noisy version of the corpus, the results for Wav2Vec2.0 are reduced by 20% while the results for Whisper are only reduced by 3%. This fact again indicates the capability of Whisper to accurately process speech recordings in non-controlled acoustic conditions.

The last covered experiment involved a pilot study on the use of FL to train ASR systems. The results indicated that an ASR trained in a federated way maintains and in some cases outperforms the performance of individual ASRs trained in a centralized

 doi:10.20944/preprints202212.0426.v1

manner by each LEA. In addition to the performance, the most important aspect of FL is that the ASR training does not involve any data sharing among LEAs, since only updates of the network parameters are transferred to a central server in charge of aggregating the model. These results are indicators about the potential use of FL to obtain a joint (and potentially richer) model combining sources of data that could not be otherwise combined. Although the benefits of using FL, it is important to consider external factors that may degrade the performance and reliability of the system. For instance, there is evidence about FL attacks that are able to retrieve speaker information from the transferred weights [60] or data poisoning attacks inside LEAs server. Different strategies can be considered to mitigate this this type of attacks such as the use of differential privacy algorithms [61] or the use of trusted execution environments.

7. Conclusions

This paper proposed the use of speech recognition and keyword spotting technologies to be applied in forensic scenarios, particularly in child exploitation domains. The aim is to provide LEAs with technology to detect the presence of offensive online audiovisual material related to child abuse. State-of-the art ASR systems based on Wav2Vec2.0 and Whisper were considered for the addressed application. The performance of both models was tested on a large set of open benchmark corpora from the literature. Therefore, the results obtained can be extended to other ASR domains. We additionally created an in-domain corpus using different open source datasets from the research community. The aim was to test the models in more realistic and operative conditions.

The ASR and KWS models were evaluated in corpora from seven Indo-European languages, including English, German, French, Spanish, Italian, Portuguese, and Polish. We obtained overall WERs ranging from 11.3% to 24.9%, depending on the language. The performance of the KWS model for the different languages ranged from 81.5% to 98.4%. The most accurate results were obtained from models trained with more data, such as English or German. The comparison between Wav2Vec2.0 and Whisper models indicated that the second one was the most accurate system in the majority of cases, especially when considering utterances in non-controlled acoustic conditions.

We also proposed a strategy for using FL to train robust ASR systems in the context of the addressed application. This is a suitable approach considering that collecting operational data from LEAs is not possible. FL approaches allow LEAs to build a common technological platform without the need to share their operational data. The results of the FL pilot indicated that similar WERs were achieved when comparing the model trained in a federated way to individual models trained in a centralized manner, even considering non-IID conditions, which has been shown to be one of the main drawbacks in FL.

For future work, the considered approaches can be extended to other forensic applications where there is a need to monitor audiovisual material from online sources. In addition, the considered technology can be combined with other speech processing methods, such as speaker and language identification, age and gender recognition, and speaker diarization. The ultimate goal is to provide LEAs with accurate tools to monitor audio from online sources, allowing them to respond in a practical and timely manner.

Author Contributions: "Conceptualization, J.C.V., A.A., and .; methodology, J.C.V. and A.A; software, J.C.V.; validation, J.C.V.; formal analysis, J.C.V. and A.A; investigation, J.C.V, A.A., and .; resources, .; data curation, J.C.V.; writing—original draft preparation, J.C.V.; writing—review and editing, J.C.V., A.A., and .; visualization, J.C.V. All authors have read and agreed to the published version of the manuscript."

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under project GRACE, grant agreement No 883341

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of the GRACE consortium (protocol code XXX and date of approval).

Data Availability Statement: All data considered in this study come from open repositories under Creative common licenses.

Conflicts of Interest: "The authors declare no conflict of interest."

Abbreviations

ASR	Automatic Speech Recognition
CTC	Connectionist Temporal Classification
FL	Federated Learning
KWS	Keyword Spotting
LEA	Law Enforcement Agency
MLS	Multilingual Librispeech
SWC	Spoken Wikipedia Corpus
TPR	True Positive Rate

References

1. Negrão, M.; Domingues, P. SpeechToText: An open-source software for automatic detection and transcription of voice recordings in digital forensics. *Forensic Science International: Digital Investigation* **2021**, *38*, 301223.
2. Alghowinem, S. A safer youtube kids: An extra layer of content filtering using automated multimodal analysis. In Proceedings of the Proceedings of SAI Intelligent Systems Conference. Springer, 2018, pp. 294–308.
3. Mariconti, E.; Suarez-Tangil, G.; Blackburn, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Serrano, J.L.; Stringhini, G. " You Know What to Do" Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *Proceedings of the ACM on Human-Computer Interaction* **2019**, *3*, 1–21.
4. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In Proceedings of the International conference on machine learning. PMLR, 2016, pp. 173–182.
5. Graves, A.; Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In Proceedings of the International conference on machine learning. PMLR, 2014, pp. 1764–1772.
6. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the ICASSP. IEEE, 2016, pp. 4960–4964.
7. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. *Advances in neural information processing systems* **2015**, *28*.
8. Lu, L.; Zhang, X.; Renais, S. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. In Proceedings of the ICASSP. IEEE, 2016, pp. 5060–5064.
9. Yao, Z.; Wu, D.; Wang, X.; Zhang, B.; Yu, F.; Yang, C.; Peng, Z.; Chen, X.; Xie, L.; Lei, X. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv preprint arXiv:2102.01547* **2021**.
10. Kriman, S.; Beliaev, S.; Ginsburg, B.; Huang, J.; Kuchaiev, O.; Lavrukhin, V.; Leary, R.; Li, J.; Zhang, Y. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In Proceedings of the ICASSP. IEEE, 2020, pp. 6124–6128.
11. Bermuth, D.; Poeppel, A.; Reif, W. Scribosermo: Fast Speech-to-Text models for German and other Languages. *arXiv preprint arXiv:2110.07982* **2021**.
12. Kolobov, R.; Okhapkina, O.; Omelchishina, O.; Platunov, A.; Bedyakin, R.; Moshkin, V.; Menshikov, D.; Mikhaylovskiy, N. Mediaspeech: Multilanguage asr benchmark and dataset. *arXiv preprint arXiv:2103.16193* **2021**.
13. Majumdar, S.; Balam, J.; Hrinchuk, O.; Lavrukhin, V.; Noroozi, V.; Ginsburg, B. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. *arXiv preprint arXiv:2104.01721* **2021**.
14. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
15. Graves, A. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711* **2012**.
16. Zhou, W.; Zheng, Z.; Schlüter, R.; Ney, H. On language model integration for rnn transducer based speech recognition. In Proceedings of the ICASSP. IEEE, 2022, pp. 8407–8411.
17. Baevski, A.; et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Sysrecognitiontems* **2020**, *33*, 12449–12460.
18. Pham, N.Q.; Waibel, A.; Niehues, J. Adaptive multilingual speech recognition with pretrained models. In Proceedings of the INTERSPEECH, 2022, pp. 3879–3883. https://doi.org/10.21437/Interspeech.2022-872.
19. Krabbenhöft, H.N.; Barth, E. TEVR: Improving Speech Recognition by Token Entropy Variance Reduction. *arXiv preprint arXiv:2206.12693* **2022**.
20. Junior, A.C.; Casanova, E.; Soares, A.; de Oliveira, F.S.; Oliveira, L.; Junior, R.C.F.; da Silva, D.P.P.; Fayet, F.G.; Carlotto, B.B.; Gris, L.R.S.; et al. CORAA: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. *arXiv preprint arXiv:2110.15731* **2021**.

21. Hsu, W.N.; Tsai, Y.H.H.; Bolte, B.; Salakhutdinov, R.; Mohamed, A. HuBERT: How much can a bad teacher benefit ASR pre-training? In Proceedings of the ICASSP. IEEE, 2021, pp. 6533–6537.
22. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2021**, *29*, 3451–3460.
23. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented Transformer for Speech Recognition. In Proceedings of the Proc. Interspeech 2020, 2020, pp. 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015.
24. Guo, P.; Boyer, F.; Chang, X.; Hayashi, T.; Higuchi, Y.; Inaguma, H.; Kamo, N.; Li, C.; Garcia-Romero, D.; Shi, J.; et al. Recent developments on espnet toolkit boosted by conformer. In Proceedings of the ICASSP. IEEE, 2021, pp. 5874–5878.
25. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. Technical report, OpenAI, 2022.
26. Voigt, P.; Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* **2017**, *10*, 10–5555.
27. Konečnỳ, J.; McMahan, H.B.; Yu, F.X.; Richtárik, P.; Suresh, A.T.; Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* **2016**.
28. Yang, Q.; Liu, Y.; Cheng, Y.; Kang, Y.; Chen, T.; Yu, H. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **2019**, *13*, 1–207.
29. Li, L.; Fan, Y.; Tse, M.; Lin, K.Y. A review of applications in federated learning. *Computers & Industrial Engineering* **2020**, *149*, 106854.
30. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine* **2020**, *37*, 50–60.
31. Dimitriadis, D.; Kumatani, K.; Gmyr, R.; Gaur, Y.; Eskimez, S.E. A Federated Approach in Training Acoustic Models. In Proceedings of the Interspeech, 2020, pp. 981–985.
32. Cui, X.; Lu, S.; Kingsbury, B. Federated acoustic modeling for automatic speech recognition. In Proceedings of the ICASSP. IEEE, 2021, pp. 6748–6752.
33. Guliani, D.; Beaufays, F.; Motta, G. Training speech recognition models with federated learning: A quality/cost framework. In Proceedings of the ICASSP. IEEE, 2021, pp. 3080–3084.
34. Hard, A.; Partridge, K.; Nguyen, C.; Subrahmanya, N.; Shah, A.; Zhu, P.; Moreno, I.L.; Mathews, R. Training Keyword Spotting Models on Non-IID Data with Federated Learning. In Proceedings of the INTERSPEECH, 2020, pp. 4343–4347. https://doi.org/10.21437/Interspeech.2020-3023.
35. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979* **2020**.
36. Wang, C.; Riviere, M.; Lee, A.; Wu, A.; Talnikar, C.; Haziza, D.; Williamson, M.; Pino, J.; Dupoux, E. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 993–1003.
37. Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G.; Collobert, R. MLS: A Large-Scale Multilingual Dataset for Speech Research. In Proceedings of the INTERSPEECH, 2020.
38. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. In Proceedings of the Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.
39. Valk, J.; Alumäe, T. VoxLingua107: a dataset for spoken language recognition. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 652–658.
40. Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; von Platen, P.; Saraf, Y.; Pino, J.; et al. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In Proceedings of the INTERSPEECH, 2022, pp. 2278–2282. https://doi.org/10.21437/Interspeech.2022-143.
41. Baumann, T.; Köhn, A.; Hennig, F. The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation* **2019**, *53*, 303–329.
42. Salesky, E.; Wiesner, M.; Bremerman, J.; Cattoni, R.; Negri, M.; Turchi, M.; Oard, D.W.; Post, M. The Multilingual TEDx Corpus for Speech Recognition and Translation. In Proceedings of the INTERSPEECH, 2021, pp. 3655–3659. https://doi.org/10.21437/Interspeech.2021-11.
43. Rousseau, A.; Deléglise, P.; Esteve, Y.; et al. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In Proceedings of the LREC, 2014, pp. 3935–3939.
44. Mirkin, S.; Jacovi, M.; Lavee, T.; Kuo, H.K.; Thomas, S.; Sager, L.; Kotlerman, L.; Venezian, E.; Slonim, N. A Recorded Debating Dataset. In Proceedings of the LREC, 2017, pp. 250–254.
45. Ogrodniczuk, M. Polish parliamentary corpus. In Proceedings of the LREC, 2018, pp. 15–19.
46. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; De Prada, A. Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media. *Applied sciences* **2019**, *9*, 5412.

 doi:10.20944/preprints202212.0426.v1

47. ECPAT. Barriers to Compensation for Child Victims of Sexual Exploitation A discussion paper based on a comparative legal study of selected countries. *ECPAT Internaltional* **2021**.
48. Richards, K. Misperceptions about child sex offenders. *Trends and issues in crime and criminal justice* **2011**, pp. 1–8.
49. EUROPOL. Online sexual coercion and extortion as a form of crime affecting children. *European Union Agency for Law Enforcement Cooperation* **2017**.
50. EUROPOL. Internet Organised Crime Threat Assessment. *European Union Agency for Law Enforcement Cooperation* **2019**.
51. EUROPOL. Exploting Isolation: Offenders and victims of online child sexual abuse during the COVID-19 pandemic. *European Union Agency for Law Enforcement Cooperation* **2020**.
52. Richey, C.; Barrios, M.A.; Armstrong, Z.; Bartels, C.; Franco, H.; Graciarena, M.; Lawson, A.; Nandwana, M.K.; Stauffer, A.; van Hout, J.; et al. Voices Obscured in Complex Environmental Settings (VOiCES) Corpus. In Proceedings of the INTERSPEECH, 2018, pp. 1566–1570. https://doi.org/10.21437/Interspeech.2018-1454.
53. Moffitt, J. Ogg Vorbis—open, free audio—set your media free. *Linux journal* **2001**, *2001*, 9–es.
54. Marcacini, R.M.; Candido Junior, A.; Casanova, E. Overview of the Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese (SE&R) Shared-tasks at PROPOR 2022. In Proceedings of the PROPOR, 2022.
55. Bai, J.; Li, B.; Zhang, Y.; Bapna, A.; Siddhartha, N.; Sim, K.C.; Sainath, T.N. Joint unsupervised and supervised training for multilingual asr. In Proceedings of the ICASSP. IEEE, 2022, pp. 6402–6406.
56. Zheng, H.; Peng, W.; Ou, Z.; Zhang, J. Advancing CTC-CRF Based End-to-End Speech Recognition with Wordpieces and Conformers. *arXiv preprint arXiv:2107.03007* **2021**.
57. Stefanel Gris, L.R.; Casanova, E.; Oliveira, F.S.d.; Silva Soares, A.d.; Candido Junior, A. Brazilian Portuguese Speech Recognition Using Wav2vec 2.0. In Proceedings of the International Conference on Computational Processing of the Portuguese Language. Springer, 2022, pp. 333–343.
58. Keshet, J.; Grangier, D.; Bengio, S. Discriminative keyword spotting. *Speech Communication* **2009**, *51*, 317–329.
59. Lengerich, C.; Hannun, A. An end-to-end architecture for keyword spotting and voice activity detection. *arXiv preprint arXiv:1611.09405* **2016**.
60. Tomashenko, N.; Mdhaffar, S.; Tommasi, M.; Estève, Y.; Bonastre, J.F. Privacy attacks for automatic speech recognition acoustic models in a federated learning framework. In Proceedings of the ICASSP. IEEE, 2022, pp. 6972–6976.
61. Geyer, R.C.; Klein, T.; Nabi, M. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557* **2017**.