University of Kent

Kent Academic Repository

McHugh, Callum, Yuan, Haiyue, Pont, Jamie, Franqueira, Virginia N. L., Arief, Budi and Hernandez-Castro, Julio C. (2026) *AGenCi: Age and gender audio classification for forensic investigations of child sexual abuse and exploitation.* In: Availability, Reliability and Security. ARES 2026 International Workshops. Lecture Notes in Computer Science . pp. 224-242. Springer Nature, Switzerland ISBN 978-3-032-35585 E-ISBN 978-3-032-35586-7.

Downloaded from

https://kar.kent.ac.uk/115911/ The University of Kent's Academic Repository KAR

The version of record is available from

https://doi.org/10.1007/978-3-032-35586-7_13

This document version

Author's Accepted Manuscript

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

AGenCi: Age and Gender Audio Classification for Forensic Investigations of Child Sexual Abuse and Exploitation

Callum McHugh[1 [0009−0008−0611−1971]], Haiyue Yuan[1 [0000−0001−6084−6719]], Jamie Pont[1 [0000−0003−0969−2464]], Virginia N. L. Franqueira[1 [0000−0003−1332−9115]✉], Budi Arief[1 [0000−0002−1830−1587]✉], and Julio Hernandez-Castro[2 [0000−0002−6432−5328]]

[1] School of Computing, University of Kent, UK
[2] Computer Systems Department, Universidad Politécnica de Madrid, Spain

Abstract. Age and gender classification using speech signals is a critical task in digital forensics, with many important applications including investigations of child sexual abuse and exploitation (CSAE). While a large amount of research has explored various machine learning (ML) and deep learning (DL) techniques for this, challenges remain in accurately identifying children's speech, especially in the context of detecting child sexual abuse material (CSAM). This is largely due to the lack of reliable/balanced datasets and the variability of children's speech data. In this study, we present a framework based on OpenAI's Whisper-medium model, integrating a custom multi-layer feedforward classification network to perform binary gender classification, binary age classification, and multi-class age classification. We call this framework AGenCi (Age and Gender audio Classification for forensic Investigations of CSAE). To fill the research gap and facilitate evaluation, three datasets are curated using well-known and publicly available speech corpora to address biases related to age, gender, and linguistic diversity. Several traditional ML models and recent DL models are selected in our benchmarking experiments. The results indicate that our framework achieves the highest accuracy of 95.6%, 94.5%, and 87.6% for the binary gender, binary age, and multi-age classification, respectively, consistently outperforming other models.

Keywords: Digital Forensics, Audio, Classification, Children, CSAE, Age, Gender, Investigation

1 Introduction

Child sexual abuse and exploitation (CSAE) represents one of the most notorious forms of violence against children, with perpetrators increasingly using digital platforms to commit these offences [26]. The volume of online child sexual abuse material (CSAM) continues to grow. According to the 2024 CyberTipline report

[27], the National Center for Missing & Exploited Children (NCMEC) in the U.S. received 29.2 million reports of suspected child sexual exploitation, including 62.9 million (i.e., 33 million video files, 28 million image files, and 1.9 million files in other formats) CSAM-related files. Similarly, the UK's Internet Watch Foundation (IWF) removed 132,700 CSAM-containing web pages, a 26% increase from the 2024, with 39% involving children under 11 years old [41]. Globally, INTERPOL's International Child Sexual Exploitation (ICSE) database reported 60% of analysed materials involved victims as young as infants, while 84% contained explicit sexual activity [17].

Within this challenging landscape, recent technological advancements are essential to support CSAE investigations. A report jointly published by the UK Department for International Trade and the Department for Science, Innovation & Technology in 2023 [36] provides an overview of the products and services developed by UK-headquartered online safety technology providers. The report presents the widespread adoption of Artificial Intelligence (AI) and Machine Learning (ML) in developing advanced tools that can provide various services, such as identifying and removing online CSAM, detecting online grooming activities, and network filtering to combat misinformation. In CSAE investigations, law enforcement agencies (LEAs) frequently encounter multimedia evidence, including images, videos, and audio recordings containing perpetrator and victim speech data [10]. This creates a need for advanced audio analysis to identify speakers and especially distinguish between adults and children.

Audio analysis can support CSAE investigation by estimating age and gender from various sources such as video, livestream, and Voice over Internet Protocol (VoIP) communications, complementing visual analysis when visual evidence is unclear or incomplete. It can also assist triage by identifying the presence of adults together with children from a large volumes of multimedia content, thereby improving intelligence generation, evidence collection, and investigative efficiency. Speech signal-based age and gender estimation and classification using both traditional ML – i.e., Decision Trees (DT), Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and Random Forests (RF) – and Deep Learning (DL) techniques have been extensively studied. Although various ML based approaches exhibit strong performance across different age and gender recognition tasks [24,37], limitations in capturing complex patterns have led to a shift towards using more sophisticated DL based approaches. By leveraging features such as Mel-Spectrograms (MEL) and Mel-Frequency Cepstral Coefficients (MFCCs) [40], DL architectures, particularly Convolutional Neural Networks (CNNs) and hybrid models have consistently outperformed traditional ML-based methods in age and gender estimation and classification tasks [1].

Nevertheless, limited work has specifically focused on exploring children's speech data [34,32], largely due to the scarcity of publicly available datasets and the variability in children's voice samples caused by linguistic development and puberty [32]. Although transformer-based architectures have demonstrated strong capabilities in pattern recognition across text, image, and video processing tasks [15]., their application to speech-signal processing, especially for age and

gender estimation and classification, remains under-explored. To address these challenges, we propose *AGenCi (Age and Gender audio Classification for forensic Investigations of CSAE)*, a framework built on OpenAI's Whisper-medium model [31][1] and enhanced with a five-layer feedforward classification network. The framework supports three tasks: **(1)** binary gender classification (male or female); **(2)** binary age classification (under 20 years old, or 20 years and older); and **(3)** multi-class age classification (under 20, 20–39, 40–59, or over 60).

Furthermore, to address biases related to age, gender and linguistic variation in training datasets, we curate dedicated datasets for these classification tasks, using a combination of the Common Voice Corpus [3], the CSLU Kids' Corpus [38], and MyST Children's Speech Corpus [30]. To comprehensively evaluate the effectiveness and accuracy of our framework, benchmarking experiments were conducted to compare it with a number of selected traditional ML and DL models. Results have shown that AGenCi achieves higher accuracies of 95.6%, 94.5%, and 87.6% for the task of binary gender, binary age, and multi-age classification, respectively, demonstrating its consistent superior performance over other benchmarking models[2].

The remainder of the paper is organised as follows. Section 2 introduces related work and the context of our work. Section 3 presents details of the main components of the proposed AGenCi framework. Section 4 describes the experimental details, including the datasets curation, model training and evaluation, while Section 5 compiles the results of our experiments, and comparing them against a number of selected benchmarking architectures. Section 6 discusses the implications of our work, along with the challenges, limitations, and future work. Finally, Section 7 concludes the paper.

2 Related Work

It is worth noting that age estimation typically refers to the task of explicitly predicting the exact age of a speaker, whereas age classification is about categorising the speaker's age into predefined age ranges. In addition, gender classification generally denotes classifying the speaker's gender as female or male.

Various ML techniques have been applied to age and gender estimation and classification tasks. Muoli et al. [24] systematically evaluated an array of regression models for age estimation and classification models for gender classification. This comparative study of various classic ML techniques concluded that Decision Tree Regressor (DTR) and Stochastic Gradient Descent Regressor (SGDR) based approaches work best for age estimation, while CatBoost, XGBoost, and Stochastic Gradient Descent (SGD) classifiers achieved the highest accuracy for the task of gender classification. Similarly, Shabbir et al. [37] compared SVM, kNN, RF, Linear Regression (LR), Gradient Boosting (GB), and Gaussian Naive

[1] `https://huggingface.co/openai/whisper-medium`

[2] All of the trained models used in this paper – along with the training and evaluation scripts – are available from `https://github.com/hyyuan/aluna-agenci`.

Bayes (GNB) for age and gender classification. The results show that SVM consistently achieved strong performance in both age and gender recognition tasks.

Various DL architectures have also been explored in this area of research. Based on CNN architecture, Tursunov et al. [40] proposed a system with a specially designed multi-attention module (MAM) to classify age and gender using spectrograms derived from speech signals. The integration of MAM with CNN effectively captured the spatial and temporal salient features from the input data, achieving accuracy scores of 96%, 73%, and 76% for gender, age, and age-gender classification, respectively. Differently, Sanchez et al. [35] proposed a framework based on a custom two-stage classification CRNN with a meta-classifier for the joint classification of speakers' gender and age groups. It consistently outperformed other architectures (i.e., CNN, Convolutional Recurrent Neural Network (CRNN), Convolutional Temporal Convolutional Networks (CTCN), and Temporal Convolutional Networks (TCN)) across all tasks. However, Sanchez et al. [35] also pointed out that the unbalanced age group dataset would influence the classification results, calling for the development of curated datasets dedicated to age and gender classification tasks.

Slightly different from the above approaches, Kwasny et al. [20] developed an x-vector-based Deep Neural Network (DNN) system for joint age estimation and gender classification. They proposed the incorporation of a deeper QuartzNet architecture [19] with a two-stage transfer learning scheme (i.e., pre-trained on the VoxCeleb dataset [25], followed by further training using the Common Voice dataset [3] for age and gender classification). The evaluation reported its state-of-the-art performance in age and gender estimation with Mean Absolute Error (MAE) of 5.12 and 5.29 years, and a Root Mean Square Error (RMSE) of 7.24 and 8.12 years for male and female speakers, respectively. The proposed approach also achieved a high gender classification accuracy of 99.6%.

Furthermore, Transformer-based architectures have shown promising results in speech classification tasks. In particular, Radford et al. [31] presented a new speech recognition system 'Whisper', which can achieve remarkable accuracy and robustness in zero-shot transfer settings, often outperforming fully supervised models across various benchmarks. Recent work [14] investigated a mixed fine-tuning approach using a pre-trained end-to-end Whisper model to improve automatic speech recognition (ASR) for children's speech by addressing the issues of age mismatch and speaking style mismatch. To the best of the authors' knowledge at the time of conducting the experiment, the most relevant work is the study conducted by Burkhardt et al. [8], where a pre-trained wav2vec 2.0 model [5] and custom classification layers were used to predict age and gender. In their study, the authors curated sample sets of publicly available datasets, including VoxCeleb2 [25], Common Voice [3], Timit [12], and aGender [7]. Experimental results across diverse datasets demonstrated the effectiveness of transformer-based models for age and gender estimation. However, the tendency to underestimate age highlights the need for improved transformer architectures and higher-quality children's speech datasets. The importance of dataset quality was also demonstrated by Safavi et al. [34], who evaluated multiple ap-

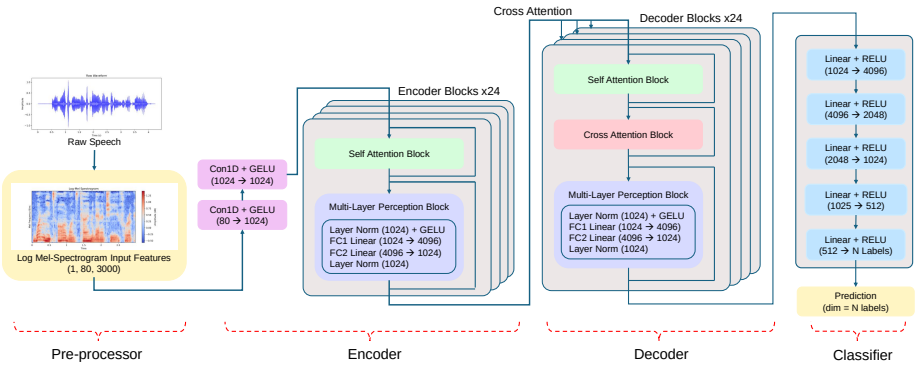


Fig. 1: The proposed AGenCi framework

proaches such as GMM-UBM (Universal Background Model), GMM-SVM, and i-vector-based approaches. Their findings showed that the classification accuracy was strongly influenced by the child's age and speech frequency, while puberty-related voice changes, particularly in boys, posed unique challenges for children's gender recognition.

More recently, Radha et al. [32] introduced the Non-Native Children's English Speech (NNCES) corpus and proposed sincNet, a CNN-based architecture, for children's age-group classification. Consistent with earlier findings [34], the authors identified several related challenges, including the lack of publicly available datasets, variability in children's speech patterns due to developing linguistic abilities and accents, and age-related vocal changes associated with growth and puberty, particularly among boys.

3 The proposed framework: AGenCi

As shown in Figure 1, the proposed AGenCi framework is mainly based on the OpenAI's Whisper-medium model [31], originally designed for Automatic Speech Recognition (ASR), and containing 769 million trainable parameters. This is a pre-trained transformer-based ASR model where audio samples are first converted to log mel-spectrograms to extract temporal and granular features of speech, then converted into embeddings that represent the context of each sample, and finally decoded to predict learned tokens. Within these classification models, the final layer of the pre-trained model was replaced with a custom classification head for the following three classification tasks: 1) binary gender classification to determine if the speaker is male or female; 2) binary age classification to classify the speaker as either below 20 years old, or 20 years and older; and 3) multi-class age classification to assign speaker to one of four age groups: $<$20, 20-39, 40-59, or $>$60.

We would like to address that previous research shows that vocal development during adolescence varies significantly, with many individuals reaching vocal

maturity after the age of 18. Relying on a strict 18-year threshold in voice-based age classification may result in misclassification [39,4], thereby to account for late or atypical vocal maturation, we use 20 years as the age threshold in this study to distinguish teenagers from adults.

3.1 Pre-processing

To ensure consistency and improve the quality of input data for classification tasks, a pre-processing pipeline is designed to transform raw speech input into feature representations (i.e., Log mel-spectrogram) for training and testing. The pipeline consists of the following steps: 1) **resampling recordings** to a standardised sample rate of 16kHz, selected as the lowest common sample rate across all self-curated datasets (Section 4.1); 2) **downmixing multi-channel audio** to mono to ensure consistency; 3) **Removing silence** using `librosa.effects.trim` [23] to eliminate low-energy segments below 20dB; 4) **Normalising scale audio amplitudes** between 0 and 1, ensuring uniform loudness levels across recordings; 5) **adding/trimming signals** to a fixed duration of 3 seconds to facilitate batch processing and model compatibility [23]; and 6) **extracting feature** using `WhisperFeatureExtractor` from Hugging Face Transformers [16], following standard preprocessing practices for the OpenAI Whisper model [43,31,16].

3.2 Encoder

Based on Whisper's architecture, feature extraction begins with two consecutive 1D convolutional (Conv1D) layers, each followed by a GELU (Gaussian Error Linear Unit) activation function. These layers extract relevant features from the log mel spectrogram input, introducing non-linearity and improving gradient flow during training, thereby enabling the model to learn complex patterns from the audio data.

As illustrated in Figure 1, the log mel spectrogram input is first processed through two Conv1D layers with GELU activation functions to generate higher-dimensional convolutional representations. These representations are subsequently passed through 24 transformer encoder blocks that progressively learn rich and contextual audio embeddings through self-attention, positional embeddings, and non-linear transformations, enabling the model to capture long-range dependencies and temporal speech patterns.

Each encoder block consists of layer normalisation (LN), a scaled dot-product self-attention mechanism, residual connections, and a feedforward multi-layer perceptron (MLP) block. The MLP component includes two fully connected (FC) layers with a GELU activation function, where the feature dimension is first expanded from 1024 to 4096 and then projected back to 1024 to enable richer feature transformations while maintaining computational efficiency. This recursive process across all 24 encoder blocks produces the final encoder representation H_{enc}.

3.3 Decoder

The decoder consists of 24 sequential transformer blocks that refine encoded representations by attending to both previously decoded outputs and the encoder representation H_{enc} through self-attention and cross-attention mechanisms. Through residual connections, layer normalisation (LN), and non-linear transformations, the decoder progressively models contextual dependencies and integrates relevant information from the encoded speech representations.

The decoder is initialised using the embedding of a special start-of-sequence (SOS) token, after which the hidden representations are updated auto-regressively at each decoding step. Similar to the encoder architecture, each decoder block contains a feedforward MLP component consisting of two FC layers with a GELU activation function between them to enable richer feature transformations. This recursive refinement process across all 24 decoder blocks produces the final decoder representation H_{dec}^{24}, which serves as the input to the downstream classification network.

3.4 Classifier

As depicted in Figure 1, the classifier is a multi-layer feedforward network that processes the decoder representation H_{dec}^{24} through four fully connected (FC) layers with rectified linear unit (ReLU) activation functions. The network progressively transforms the input representation through latent dimensions of 4096, 2048, 1024, and 512, enabling hierarchical feature compression while preserving discriminative age- and gender-related speech characteristics.

The final FC layer maps the 512-dimensional representation to an output vector of size C, where the value of C depends on the classification task: $C = 1$ for binary classification tasks (i.e., binary age and binary gender classification), whereas $C = 4$ for the multi-class age classification task. The resulting output O is subsequently passed through a sigmoid activation function for binary classification or a softmax activation function for multi-class classification.

4 Experiments

4.1 Datasets

As identified in Section 2, curated datasets are needed to improve classification accuracy and mitigate biases related to age, gender, and linguistic variation. To this end, three datasets are carefully curated in this study to ensure high-quality and well-balanced data distributions, which are essential for robust model training and evaluation. In addition, documenting the dataset curation process supports reproducibility and provides guidance for researchers and practitioners seeking to adopt similar approaches.

Selecting relevant datasets. Table 1 shows a list of (non-exhaustive) datasets that have been used in the research discussed in Section 2. After excluding non open-access datasets, those lacking children's speech data, and those that

Table 1: Summary of speech datasets for age and gender classification

Dataset	Age Range Information	Data Labels	Language	Open access
VoxCeleb [25]	Adults (over 1,000 interviews of celebrities from YouTube)	G	English	✓
Common Voice [3]	6% of samples <18 years old	G & A	English	✓
Korean Speech Recognition [40,2]	Elderly men and women	G & A	Korean	✗*
ELSDSR dataset [11]	Age 24-63	G & A	English	✗#
Child Speech, Kid Speaking [13]	Children (but no specific age information)	No G/A	Multi-language	✓
831 Hours British English Speech Data by Mobile Phone [28]	51% speakers are 13-25 years old	Unknown	English	✗
DARPA-TIMIT [33]	N/A	G & A$^{\vartheta}$	English	✓
The CMU Kids [9]	Age 6-11	Unknown	English	✗
CSLU OGI Kids [38]	Children from Kindergarten to Grade 10	None	English	✓
FARSDAT [6]	Age 10-80	G & A	Farsi	✗
aGender [7]	Age 7-80 (low amount of samples for age group 7-14)	G$^{\alpha}$ & A	German	✗
My Science Tutor (MyST) [30]	Children from 3rd, 4th, 5th Grade	A$^{\beta}$	English	✓
Non-Native Children English Speech (NNCES) Corpus [32]	Age 8-12 (all are non-native English speakers)	G & A	English$^{\omega}$	✓

G = Gender, A = Age, $^{\vartheta}$This information is included in the dataset metadata, $^{\alpha}$Child gender is not labelled, $^{\beta}$In the format of '(School) Grade label', *Only Koreans can apply for data, $^{\#}$Data can be acquired upon request, but the link is broken, $^{\omega}$English as a second language of all speakers.

are not in English, the following datasets were selected for further curation: 1) **The Common Voice Corpus** [3] contains large quantities (i.e., 2,586 hours of validated audio from 90,474 English speakers) of age (e.g., Teens, Twenties, Thirties,...) and gender-labelled (i.e., Male, Female or Other) data and is commonly used for age and gender classification tasks. 2) **The CSLU Kids' Corpus** [38] is a dataset containing samples of speech from 1,100 children from the U.S. between the school years Kindergarten to Grade 10 (approximately ages 4-16). 3) **MyST Children's Speech Corpus** [30] contains collections of 400 hours of children's conversational speech, including 230K utterances from approximately 1.3K third/fourth/fifth-grade students across 10.5K virtual tutor sessions.

Curating datasets. For consistency and clarity, the Common Voice Corpus, the CSLU Kid's Corpus, and MyST Children' Speech Corpus are referred to as Dataset 1 (D1), Dataset 2 (D2), and Dataset 3 (D3), respectively, for the remainder of this paper. In D1, only samples that were between 2 and 8 seconds were extracted using the `duration` label, and any samples which did not include an age or gender label were removed. Participants with a gender labelled as `Other` were also omitted due to their low frequency, ensuring a balanced representation of the remaining gender classes. To facilitate multi-class classification, age categories were aggregated into broader groups. We use age group label `0`, `1`, `2`, and `3` to represent *Teens* (<20), *Twenties and Thirties*, *Forties and Fifties*, and *Sixties and Above*, respectively. This accounts for the non-linear nature of vocal development, recognising that speech patterns do not change uniformly within each decade of life. These larger groupings, therefore, create a generalised view of vocal patterns and features. Similarly, gender labels were mapped to binary values (female = 0 and male = 1).

However, a key limitation of D1 is the absence of speech samples from participants younger than the Teens category, highlighting the need for additional data to adequately represent the age group `0` class. To address this gap, D2 is selected as it provides speech samples from children aged 4 to 16. One main limitation of D2 is the absence of gender labels. Prior studies have highlighted that young children exhibit minimal differences in vocal features between genders, making gender classification more challenging. To the best of our knowledge, no publicly available English-language child speech datasets include gender annotations.

Considering the good coverage of speech samples covering school-aged children from D2, it is useful and necessary to manually label samples from this dataset. D2 is structured into folders corresponding to school year groups (i.e., K, 1, 2, 3, ..., 10), with individual subfolders for each participant. As only a subset of the dataset was needed to represent the age group `0` class, a systematic protocol was employed to ensure reliable manual annotations: (a) approximately 3–5 audio samples per participant were randomly selected and subsequently evaluated and assigned a gender label (i.e., could be female or male) by the first author; (b) only participants whose assigned gender label remained consistent across all selected 3-5 audio samples were considered for labelling; and (c) only samples with annotated labels and "Good" quality scores were kept (please refer to [38] for more information about the quality score).

The final labels followed the same binary mapping as D1 (i.e., female = `0`, male = `1`), ensuring consistency across datasets. Overall, a total of 550 participants were annotated, with 25 male and 25 female participants selected from each of the 11 age groups. We also noticed that D2 contains some low-quality or noisy recordings, which may negatively impact model performance, emphasising the need for high-quality data to better represent age group `0`.

To mitigate such potential risk, D3 is incorporated as it is a more recent dataset with cleaner, higher-quality speech samples. The inclusion of D3 enhances the balance of the curated datasets by supplementing age group `0` with additional reliable data. In a nutshell, both D2 and D3 contributed to improving the representation of younger children, ensuring a more comprehensive and balanced dataset for the downstream classification tasks.

Overall, we mixed D1, D2, and D3 to curate three distinct datasets for separate classification tasks, namely the multi-class age dataset, the binary age dataset, and the binary gender dataset, where each curated dataset has equal representation of samples for both age and gender. The demographic distributions of these three curated datasets are presented in Table 2, Table 3(a), and Table 3(b), with training, validation, and testing splits kept as close to 70:15:15 as possible. It is worth noting that the split is speaker-independent, meaning that samples from the same speaker do not appear in multiple splits. Additionally, within all datasets, the age group `0` class maintains an approximately 3:1 ratio of child data (sourced from D2 and D3) to Teens' data (sourced from D1), ensuring a greater proportion of child speech samples to enhance inter-class differentiation. Due to the licensing restrictions, D1, D2, and D3 cannot be publicly released, although they are available upon request.

Table 2: Multi-class age dataset composition

Training

Dataset (D)	0 (<20)	1 (20-39)	2 (40-59)	3 (≥60)	Total
D3	7,500	–	–	–	7,500
D1 Male	1,250	5,000	5,000	5,000	16,250
D1 Female	1,250	5,000	5,000	5,000	16,250
Total	*10,000*	*10,000*	*10,000*	*10,000*	*40,000*

Validation

Dataset (D)	0 (<20)	1 (20-39)	2 (40-59)	3 (≥60)	Total
D3	1,875	–	–	–	1,875
D1 Male	312	1,250	1,250	1,250	4,062
D1 Female	313	1,250	1,250	1,250	4,063
Total	*2,500*	*2,500*	*2,500*	*2,500*	*10,000*

Testing

Dataset (D)	0 (<20)	1 (20-39)	2 (40-59)	3 (≥60)	Total
D3	1,875	–	–	–	1,875
D1 Male	313	1,250	1,250	1,250	4,063
D1 Female	312	1,250	1,250	1,250	4,062
Total	*2,500*	*2,500*	*2,500*	*2,500*	*10,000*

Table 3: Dataset composition for binary classification tasks

Training

Dataset (D)	0 (<20)	1 (≥20)	Total
D3	15,000	–	15,000
D1 Male	2,500	9,999	12,499
D1 Female	2,500	9,999	12,499
Total	*20,000*	*19,998*	*39,998*

Validation

Dataset (D)	0 (<20)	1 (≥20)	Total
D3	3,750	–	3,750
D1 Male	625	2,499	3,124
D1 Female	625	2,499	3,124
Total	*5,000*	*4,998*	*9,998*

Testing

Dataset (D)	0 (<20)	1 (≥20)	Total
D3	3,750	0	3,750
D1 Male	625	2,502	3,127
D1 Female	625	2,502	3,127
Total	*5,000*	*5,004*	*10,004*

(a) Age

Training

Dataset (D)	0 (Female)	1 (Male)	Total
D2 Male	–	3,286	3,286
D2 Female	3,280	–	3,280
D1 Male	–	13,990	13,990
D1 Female	13,991	–	13,991
Total	*17,271*	*17,276*	*34,547*

Validation

Dataset (D)	0 (Female)	1 (Male)	Total
D2 Male	–	681	681
D2 Female	705	–	705
D1 Male	–	2,999	2,999
D1 Female	3,001	–	3,001
Total	*3,706*	*3,680*	*7,386*

Testing

Dataset (D)	0 (Female)	1 (Male)	Total
D2 Male	–	695	695
D2 Female	712	–	712
D1 Male	–	2,998	2,998
D1 Female	2,998	–	2,998
Total	*3,710*	*3,693*	*7,403*

(b) Gender

4.2 Training and evaluation

For each classification task, a dedicated model is fine-tuned on a specifically curated dataset comprising labelled audio samples. The details of separate training and evaluation datasets used for each classification task are provided in Table 2 and Table 3. These datasets are designed to include diverse representations of gender, age, and English accents, aiming to mitigate biases and enhance the generalisation capability of the models.

The fine-tuning process follows standard ML procedures implemented using the PyTorch library [29]. Each dataset is pre-processed and loaded into `DataLoaders` for efficient batch processing on a GPU, with batch sizes of 20

for training and 32 for validation and testing. The Whisper model is fine-tuned for 30 epochs using both Adam [18] and AdamW optimisers [22], with learning rates (LR) set to 1E-5 and 2E-5, respectively. All other optimiser parameters are retained at their default values. This configuration enables the model to update its weights iteratively, refining its ability to capture speech features relevant to each classification task. The loss functions and evaluation metrics were tailored to each classification task, where `BCEWithLogitsLoss()` was used for the binary classification task, including a `Sigmoid()` activation function mapping the output of the logit function to probability scores, indicating the likelihood of predicting the positive class (`1`). For multi-class classification, `CrossEntropyLoss()` function was used, including a `Softmax()` activation layer, transforming the logit output into a probability distribution across multiple classes.

Models were trained on a secure high-performance computing (HPC) server equipped with multiple GPU partitions featuring different hardware configurations. The Ampere partition consisted of four NVIDIA A100 80GB GPUs, dual Intel Gold 5317 CPUs, and 348GB RAM. The Volta partition utilised one NVIDIA TITAN V GPU with dual Intel Xeon E5-2620 CPUs and 144GB RAM, while the Pascal partition provided two NVIDIA Tesla P100 GPUs, three Intel Gold 6136 CPUs, and 256GB RAM. The final models were trained on the Pascal partition using multiple GPUs for parallel training to reduce overall training time. We empirically evaluated our methods using the curated datasets compared with a number of selected architectures ranging from traditional classification architectures, including kNN, SVC, Linear SVC, and RF, to more recent DL-based architectures, including CNN [44], CRNN [35], CRNN+TCN [35], ResNet50 [1], and wav2vec 2.0 [8].

The selected DL-based models were chosen to represent a diverse range of DL architectures, from the baseline CNN to more advanced variants, such as CRNN and ResNet50, and finally to the transformer-based wav2vec 2.0, which is particularly relevant to the task. It is also worth noting that architectures/pre-trained models used for benchmarking have only been tested for age estimation, age classification, or gender classification in their associated studies. To adapt them for the three classification tasks in this study, minor modifications (such as adjusting the final output layer) were applied to ensure all models have consistent output for systematic comparison and analysis. The results have been evaluated using four primary metrics: accuracy, precision, recall, and F1-score for binary classification tasks (binary gender and binary age) and their macro-averaged counterparts for multi-age classification, aiming to provide a comprehensive assessment and comparison of the performance across various models.

5 Results

As shown in Table 4, for the binary gender classification task, both the AGenCi model and wav2vec 2.0 achieve the highest accuracy of 0.956, significantly outperforming other benchmarking models. By looking at the precision and recall scores, the AGenCi model gets a precision score of 0.954 and a recall score of

Table 4: Performance comparison across different models for binary gender classification, binary age classification, and multi-age classification.

	Binary Gender				**Binary Age**				**Multi-Age Classification**			
Model	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision M	Recall M	F1 M
kNN	0.649	0.692	0.533	0.602	0.727	0.685	0.839	0.755	0.381	0.374	0.152	0.349
Linear SVC	0.802	0.803	0.800	0.801	0.738	0.733	0.749	0.741	0.412	0.403	0.388	0.412
RF	0.851	0.847	0.855	0.851	0.861	0.921	0.790	0.851	0.546	0.529	0.566	0.550
SVC	0.883	0.868	0.903	0.885	0.866	0.920	0.803	0.857	0.573	0.572	0.604	0.576
ResNet 50 [1]	0.826	0.831	0.818	0.825	0.825	0.828	0.820	0.824	0.519	0.485	0.544	0.520
CRNN [35]	0.844	0.869	0.811	0.839	0.833	0.852	0.805	0.828	0.578	0.657	0.539	0.586
CNN+TCN [35]	0.899	0.889	0.913	0.900	0.858	0.883	0.825	0.853	0.591	0.592	0.582	0.654
CNN [44]	0.929	0.924	0.935	0.930	0.894	0.909	0.876	0.892	0.710	0.826	0.721	0.715
wav2vec 2.0 [8]	**0.956**	0.951	**0.962**	**0.957**	0.933	0.950	0.914	0.932	0.840	0.898	0.917	0.841
AGenCi	**0.956**	**0.954**	0.958	0.956	**0.945**	**0.966**	**0.923**	**0.944**	**0.876**	**0.928**	**0.935**	**0.877**

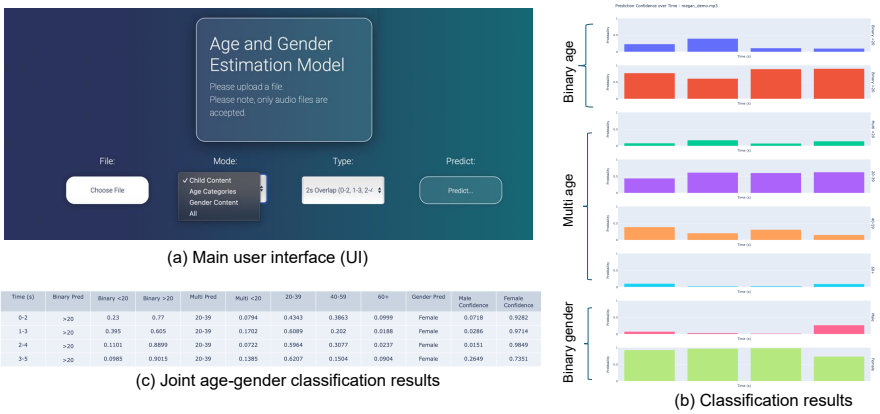
0.958, while wav2vec 2.0 also shows similar performance with a precision score of 0.951 and a recall score of 0.962, respectively.

This indicates that both models are capable of offering a low false positive rate and a high true positive rate. The F1-scores further confirm the robustness and superiority of both transformer-based models over other benchmarking models. As compared in Table 4, wav2vec 2.0 outperforms the AGenCi model only by a very small margin in both recall and F1-score.

For the binary age classification task, the AGenCi model achieves the highest accuracy of 0.945, while wav2vec 2.0 and CNN also exhibit great accuracy with 0.933 and 0.894, respectively. Regarding the precision and recall scores, the AGenCi model achieves a score of 0.966 and a score of 0.923, respectively, indicating that it outperforms other models in reducing false positives while offering a high detection rate of true positives. This is essential in the context of CSAM investigation, because missing a true positive is more costly than a false alarm. The F1-score also aligns with these findings, where the AGenCi model produces the highest score of 0.944.

The multi-age classification task appears to be more challenging, as indicated by the overall lower metric values compared to the binary gender and binary age classification results. Nevertheless, the AGenCi model outperforms other models with the highest accuracy of 0.876. The AGenCi model achieves the highest macro-averaged precision score of 0.928 and macro-averaged recall of 0.935, demonstrating its capability to correctly identify different age groups while maintaining minimal false positives. The macro-averaged F1-score represents the combined measure of class-wise precision and recall. Again, the AGenCi model achieves the best score of 0.935, proving its overall superiority in this classification task compared with other models.

Overall, the various metrics used in the evaluation indicate that transformer-based models, including wav2vec 2.0 and the AGenCi model, consistently outperform other models across all three classification tasks. More specifically, wav2vec 2.0 is slightly better than the AGenCi model in the task of binary gender classification, while the AGenCi model is consistently better than wav2vec 2.0 in both binary age classification and multi-age classification tasks.



(a) Main user interface (UI)

(b) Classification results

Time (s)	Binary Pred	Binary <20	Binary >20	Multi Pred	Multi <20	20-39	40-59	60+	Gender Pred	Male Confidence	Female Confidence
0-2	>20	0.23	0.77	20-39	0.0794	0.4343	0.3863	0.0999	Female	0.0718	0.9282
1-3	>20	0.395	0.605	20-39	0.1702	0.6089	0.202	0.0188	Female	0.0286	0.9714
2-4	>20	0.1101	0.8899	20-39	0.0722	0.5964	0.3077	0.0237	Female	0.0151	0.9849
3-5	>20	0.0985	0.9015	20-39	0.1385	0.6207	0.1504	0.0904	Female	0.2649	0.7351

(c) Joint age-gender classification results

Fig. 2: Screenshots of the developed tool for age and gender estimation: (a) Main UI, (b) UI for classification results, and (c) UI for joint age-gender classification.

5.1 Prototype tool for LEAs

To evaluate the effectiveness and accuracy of the proposed framework further, we designed and developed a prototype tool for LEAs, which has an user interface (UI) as illustrated in Figure 2(a) to allow users to configure classification settings, including file selection, sample sizes, and classification mode selection, to customise the classification process. Once the classification is completed, the tool provides results for the binary gender classification task, the binary age classification task, and the multi-age classification task as shown in Figure 2(b). Additionally, the detailed descriptive statistics for joint age-gender classification tasks are summarised in a table as illustrated in Figure 2(c).

This tool is being tested in a pilot study with LEAs to assess how practical and useful this tool is in real-world forensic investigations. This pilot study aims to gather qualitative feedback regarding its usability and efficiency, and the feedback will be used to refine the design of both the framework and the interface to ensure it aligns with the operational needs of CSAE investigations. A detailed analysis of the qualitative data will be included in a separate study as part of our future work.

6 Discussion

6.1 Key Implications

The development of AGenCi and its associated tool not only has the potential to improve operational efficiency for law enforcement but is also useful for broader applications. It could be integrated into other speech processing tasks in various contexts and support researchers and practitioners conducting similar classification tasks. As such, its implications extend well beyond law enforcement and could reach into wider research and practical contexts, such as the following:

- AGenCi could significantly speed up the identification of potential CSAM by narrowing down the subset of audio and video samples that require further investigation. Samples that are less likely to be relevant (e.g., only contain adult voices) can be temporarily deferred or removed from the suspected list, thereby streamlining investigative workflows. This could save valuable time and resources, allowing authorities to prioritise and focus their efforts more effectively.
- AGenCi could reduce the psychological burden on investigators by allowing them to identify potential CSAM based on audio samples rather than direct exposure to large amounts of video graphics. This alternative screening method could mitigate the emotional toll and long-term mental health risk associated with the repetitive viewing of distressing and offensive content.
- Going beyond the law enforcement context, AGenCi can be used for more benign or creative use cases. For instance, it could be incorporated into automated speech transcription systems to facilitate the assignment of speaker labels (e.g., "Female speaker 1", "Female speaker 2") to further enhance the readability and contextual awareness of speech transcripts. In addition, AGenCi could be utilised in voice-activated technologies, such as home assistants and voice-based authentication systems, to improve user experience and system accuracy, allowing more personalised or context-aware interactions. We wish to emphasise that we have not tested AGenCi under such broader contexts, but encourage the wider research community to explore these alternative avenues.
- By making AGenCi open access, we hope to encourage broader adoption in both academic and applied settings, where researchers and practitioners could adapt or extend it for various purposes including, but not limited to, child protection, accessibility technologies, and socio-linguistic studies.

6.2 Limitations and Future Work

One limitation of our work is the lack of inclusivity in the training datasets. For example, variations in language, such as accents and slang, are often underrepresented. This could cause potential biases and reduced accuracy for recognising speakers from diverse linguistic backgrounds, not to mention that the variations in children's speech could bring yet more complexity. In addition, going beyond the scope of this study, the absence of speech data from non-binary individuals could affect the effectiveness of gender classification models, as most existing datasets rely on binary gender labels. However, previous studies indicated that the biological differences in pre-pubescent voices are less critical compared to adults, as hormonal changes during puberty (e.g., vocal cord thickening in males) create more distinct acoustic features [21]. This considered, we would argue that such limitation might have less impact on the gender classification involving children.

Another limitation is that manual gender labelling of children's speech was conducted by a single annotator. We acknowledge that this may introduce annotation bias, particularly because children's voices can be acoustically ambiguous.

However, preliminary in-the-wild testing with industry partners suggests that the current approach remains practically effective, indicating that the impact of potential annotation bias may be limited. Future work should involve multiple independent annotators and report inter-annotator agreement, such as Cohen's kappa or Krippendorff's alpha, to validate the consistency of manually assigned gender labels.

Third main limitation in the context of CSAM investigation is related to the probabilistic outcome generated by ML or DL-based age and gender classification tool, meaning that even a low false positive rate can result in significant resource demands to differentiate true positives from false positives. This inherent non-determinism could potentially lead to the unnecessary scrutiny of individuals during an investigation.

Moving forward, although curating datasets from three sources helps mitigate the limitation related to inclusivity in the training data, further work is needed to develop 1) more inclusive datasets with broader demographic representations, and 2) bias-aware models to promote fairness in age and gender classification. To align with industry best practices and regulatory obligations, we also recommend that *researchers and practitioners document their efforts to minimise such biases while designing privacy-preserving and non-discriminatory AI systems.* To align with the proposed CSAM regulations [42], AI-powered detection tools should minimise privacy intrusion while actively reducing false positive and false negative rates. We recommend that *researchers and practitioners should prioritise these factors to ensure the effectiveness, fairness, and ethical deployment of such technologies.*

Moreover, the development of AI models for CSAM detection often faces the challenge of the "black box" issue. The lack of transparency can affect the trust, explainability, and traceability of the tool. To this end, we recommend that *researchers and practitioners document datasets used in the training process, decision-making processes, and outcomes to the best possible standard, to enhance the transparency of AI-powered CSAM tools as well as to facilitate the development of explainable AI components.* Furthermore, there are challenges to deal with variability in children's speech patterns in age and gender classification tasks, as mentioned in Section 2. We believe *more research could be carried out to specifically learn how to better capture the variability and developmental characteristics of children's voice samples, such as exploring acoustic features that are sensitive to vocal tract length changes or articulation patterns.* This can potentially address the "black box" challenge mentioned above.

While the AGenCi framework demonstrates promising performance, two potential research directions could enhance its applicability in real-world scenarios. First, the proposed model is not specifically designed to deal with audio samples that have non-stationary noise, such as traffic and machinery. It is unclear how robust the model is to distinguish human speech from such noise, and to what extent this may impact classification accuracy. We recommend *further research into methods that can reliably separate speech from dynamic acoustic environments.* Second, human conversations or voices that are partially obscured by

background noise, or in some cases, imperceptible to the human ear, may have critical information, particularly in CSAE investigations, yet, this remained unexplored in this study. We recommend *that additional work is needed to enhance the detection of subtle or low-salience speech signals in noisy settings.*

7 Conclusion

This study reports our work in developing the AGenCi (Age and Gender audio Classification for forensic Investigations of CSAE) framework based on OpenAI's Whisper-medium model with a custom classification head for three distinct classification tasks: binary age classification, binary gender classification, and multi-age classification. To address the identified research gap on the lack of curated datasets for age and gender classification involving children, this study also presents the detailed methodology for curating three dedicated datasets using a combination of three existing speech corpora. To test the effectiveness and accuracy of the proposed framework, benchmarking experiments involving a number of selected traditional ML and DL models were carried out to compare their performance against our framework, using the curated datasets. The evaluation indicates that AGenCi consistently outperforms other models across all classification tasks.

Acknowledgments. This work was supported by the funding received from the EU ISF Grant Agreement No. 101084929 on "Child protection centred strategies to fight against sexual abuse and exploitation (ALUNA)", `https://www.aluna-isf.eu/`.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alnuaim, A.A., Zakariah, M., Shashidhar, C., et al.: Speaker Gender Recognition Based on Deep Neural Networks and ResNet50. Wireless Communications and Mobile Computing **2022**(1), 4444388 (2022)
2. Anvarjon: Age-Gender-Classification. `https://github.com/Anvarjon/Age-Gender-Classification/` (2024)
3. Ardila, R., Branson, M., Davis, K., et al.: Common Voice: A Massively-Multilingual Speech Corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4218–4222. European Language Resources Association (May 2020)
4. Asci, F., Costantini, G., Di Leo, P., et al.: Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender. Sensors **20**(18) (2020)
5. Baevski, A., Zhou, Y., Mohamed, A., et al.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: Proceedings of Advances in Neural Information Processing Systems. vol. 33, pp. 12449–12460 (2020)
6. Bijankhan, M., Sheikhzadegan, J., Roohani, M.R.: FARSDAT-The speech database of Farsi spoken language. In: Proceedings of Australian Conference on Speech Science and Technology. pp. 826–830 (1994)

7. Burkhardt, F., Eckert, M., Johannsen, W., Stegmann, J.: A Database of Age and Gender Annotated Telephone Speech. In: Proceedings of 7th International Conference on Language Resources and Evaluation (LREC'10). pp. 1562–1565 (2010)
8. Burkhardt, F., Wagner, J., Wierstorf, H., et al.: Speech-based Age and Gender Prediction with Transformers. In: Proceedings of Speech Communication; 15th ITG Conference. pp. 46–50 (2023)
9. Eskenazi, M., Mostow, J., Graff, D.: The CMU Kids Corpus. https://catalog.ldc.upenn.edu/LDC97S63 (1997). https://doi.org/10.35111/b4v0-ff65, Linguistic Data Consortium
10. Europol: Global crackdown on kidflix, a major child sexual exploitation platform with almost two million users. https://www.europol.europa.eu/media-press/newsroom/news/global-crackdown-kidflix-major-child-sexual-exploitation-platform-almost-two-million-users (2025)
11. Feng, L., Hansen, L.K.: A new database for speaker recognition. IMM, Informatik og Matematisk Modelling, DTU (2005)
12. Garofolo, J.S., Lamel, L.F., Fisher, W.M., et al.: TIMIT acoustic-phonetic continuous speech corpus (1993). https://doi.org/https://doi.org/10.35111/17gk-bn40, Linguistic Data Consortium
13. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., et al.: AudioSet: Child speech, kid speaking. https://research.google.com/audioset/dataset/child_speech_kid_speaking.html (2017), accessed: 2025-01-23
14. Graave, T., Li, Z., Lohrenz, T., Fingscheidt, T.: Mixed children/adult/childrenized fine-tuning for children's asr: How to reduce age mismatch and speaking style mismatch. In: Proceedings of Interspeech 2024. pp. 5188–5192 (2024)
15. Han, K., Wang, Y., Chen, H., et al.: A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 87–110 (2023)
16. Hugging Face: WhisperFeatureExtractor. https://huggingface.co/docs/transformers/en/model_doc/whisper (2016)
17. INTERPOL: International child sexual exploitation database. https://www.interpol.int/en/Crimes/Crimes-against-children/International-Child-Sexual-Exploitation-database (nd)
18. Kingma, D.P., Ba, J.: HuggingFace's Transformers: State-of-the-art Natural Language Processing (2017). https://doi.org/10.48550/arXiv.1412.6980
19. Kriman, S., Beliaev, S., Ginsburg, B., et al.: Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions. In: Procs. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 6124–6128 (2020)
20. Kwasny, D., Hemmerling, D.: Joint gender and age estimation based on speech signals using x-vectors and transfer learning (2020). https://doi.org/10.48550/arXiv.2012.01551
21. Lee, S., Potamianos, A., Narayanan, S.: Acoustics of children's speech: Developmental changes of temporal and spectral parameters. The Journal of the Acoustical Society of America **105**(3), 1455–1468 (1999)
22. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (2019). https://doi.org/10.48550/arXiv.1711.05101
23. McFee, B., Raffel, C., Liang, D., et al.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th annual Scientific Computing with Python conference. pp. 18–24. SciPy Proceedings, Austin, Texas, USA (2015)
24. Munoli, B.K., Jain, K.A.K., Kumar, P., et al.: Human Voice Analysis to Determine Age and Gender. In: Proceedings of 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC). pp. 1–4. IEEE (2023)

25. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: Voxceleb: Large-scale speaker verification in the wild. Computer Speech & Language **60**, 101027 (2020)
26. NCA: Child sexual abuse and exploitation. https://www.nationalcrimeagency.gov.uk/what-we-do/crime-threats/child-sexual-abuse-and-exploitation
27. NCMEC: Cybertipline 2024 report. https://www.missingkids.org/cybertiplinedata (2024)
28. Nexdata AI: 831 hours british english speech data by mobile phone. https://github.com/Nexdata-AI/831-Hours-British-English-Speech-Data-by-Mobile-Phone (2025), accessed: 2025-01-23
29. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
30. Pradhan, S., Cole, R.A., Ward, W.H.: My science tutor (MyST)–a large corpus of children's conversational speech. In: Procs. 2024 Joint Int.l Conf. on Comp. Linguistics, Lang. Resources and Eval. (LREC-COLING). pp. 12040–12045 (2024)
31. Radford, A., Kim, J.W., Xu, T., et al.: Robust speech recognition via large-scale weak supervision. In: Proceedings of 40th International Conference on Machine Learning. ICML'23, JMLR.org, Honolulu, Hawaii, USA (2023)
32. Radha, K., et al.: Automatic speaker and age identification of children from raw speech using sincNet over ERB scale. Speech Communication **159**, 103069 (2024)
33. Rémy, P., Fekadu, M.: The DARPA TIMIT acoustic-phonetic continuous speech corpus. https://github.com/philipperemy/timit (2025), accessed: 2025-01-23
34. Safavi, S., Russell, M., Jančovič, P.: Automatic speaker, age-group and gender identification from children's speech. Comp. Speech & Language **50**, 141–156 (2018)
35. Sánchez-Hevia, H.A., Gil-Pita, R., Utrilla-Manso, M., et al.: Age group classification and gender recognition from speech with temporal convolutional neural networks. Multimedia Tools and Applications **81**(3), 3535–3552 (2022)
36. Saqib Bhatti MP, Lord Offord of Garvel CVO: Directory of uk safety tech providers. Tech. rep., DIT and DSIT, UK (2023), https://www.gov.uk/government/publications/directory-of-uk-safety-tech-providers
37. Shabbir, M., Hussain, A., Khan, M.M.: Age and gender estimation through speech: A comparison of various techniques. In: Proceedings of 18th International Conference on Emerging Technology (ICET). pp. 228–233. IEEE (2023)
38. Shobaki, K., Hosom, J.P., Cole, R.: CSLU: Kids' speech version 1.1 (2007). https://doi.org/10.35111/q5tn-8096, Linguistic Data Consortium
39. Taylor, S., Dromey, C., Nissen, S.L., et al.: Age-related changes in speech and voice: spectral and cepstral measures. Journal of Speech, Language, and Hearing Research **63**(3), 647–660 (2020)
40. Tursunov, A., Mustaqeem, Choeh, J.Y., et al.: Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms. Sensors **21**(17) (2021)
41. UK Gov: Online harms: interim codes of practice. https://www.gov.uk/government/publications/online-harms-interim-codes-of-practice (2020)
42. Vavoula, N., de Swart, L., Essink, J., et al.: Proposal for a regulation laying down the rules to prevent and combat child sexual abuse: Complementary impact assessment (2023), https://orbilu.uni.lu/bitstream/10993/62648/1/EPRS_STU%282023%29740248_EN.pdf
43. Wolf, T., Debut, L., et al.: HuggingFace's Transformers: State-of-the-art Natural Language Processing (2020). https://doi.org/10.48550/arXiv.1910.03771
44. Yücesoy, E.: Gender Recognition from Speech Signal Using CNN, KNN, SVM and RF. Procedia Computer Science **235**, 2251–2257 (2024), International Conference on ML and Data Eng. (ICMLDE 2023)